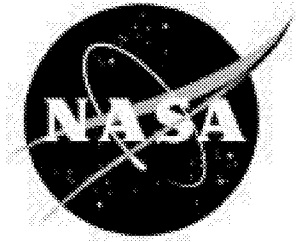


NASA/TM-2002-211638



Experimental OAI-Based Digital Library Systems

Edited by:

Michael L. Nelson

Langley Research Center, Hampton, Virginia

Kurt Maly and Mohammad Zubair

Old Dominion University, Norfolk, Virginia

Diann Rusch-Feja

Max Planck Institute for Human Development, Berlin, Germany

April 2002

The NASA STI Program Office . . . in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- **TECHNICAL PUBLICATION.** Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers, but having less stringent limitations on manuscript length and extent of graphic presentations.
- **TECHNICAL MEMORANDUM.** Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- **CONTRACTOR REPORT.** Scientific and technical findings by NASA-sponsored contractors and grantees.
- **CONFERENCE PUBLICATION.** Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.
- **SPECIAL PUBLICATION.** Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results . . . even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at ***<http://www.sti.nasa.gov>***
- Email your question via the Internet to ***help@sti.nasa.gov***
- Fax your question to the NASA STI Help Desk at (301) 621-0134
- Telephone the NASA STI Help Desk at (301) 621-0390
- Write to:
NASA STI Help Desk
NASA Center for AeroSpace Information
7121 Standard Drive
Hanover, MD 21076-1320

NASA/TM-2002-211638



Experimental OAI-Based Digital Library Systems

Edited by:

Michael L. Nelson

Langley Research Center, Hampton, Virginia

Kurt Maly and Mohammad Zubair

Old Dominion University, Norfolk, Virginia

Diann Rusch-Feja

Max Planck Institute for Human Development, Berlin, Germany

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

April 2002

Available from:

NASA Center for AeroSpace Information (CASI)
7121 Standard Drive
Hanover, MD 21076-1320
(301) 621-0390

National Technical Information Service (NTIS)
5285 Port Royal Road
Springfield, VA 22161-2171
(703) 605-6000

Experimental OAI-Based Digital Library Systems

**Workshop Held at the 5th European Conference on Research and
Advanced Technology for Digital Libraries, (ECDL 2001)**

Editors:

Michael L. Nelson
NASA Langley Research Center
Hampton, Virginia, USA
m.l.nelson@larc.nasa.gov

Kurt Maly, Mohammad Zubair
Old Dominion University
Norfolk, Virginia, USA
{maly, zubair}@cs.odu.edu

Diann Rusch-Feja
Max Planck Institute for Human Development
Berlin, Germany
ruschfeja@mpib-berlin.mpg.de

Abstract

The objective of Open Archives Initiative (OAI) is to develop a simple, lightweight framework to facilitate the discovery of content in distributed archives (<http://www.openarchives.org>). The focus of the workshop held at the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 20001) was to bring researchers in the area of digital libraries who are building OAI based systems so as to share their experiences, problems they are facing, and approaches they are taking to address them. The workshop consisted of invited talks from well-established researchers working in building OAI based digital library system along with short paper presentations.

Introduction

The Open Archives Initiative (OAI) (www.openarchives.org) is an international consortium focused on furthering the interoperability of digital libraries (DLs) through the use of "metadata harvesting". Many previous DL interoperability projects focused on "distributed searching" as the method for federating different DLs into a single service. While feasible for small numbers of nodes (e.g., < 20), large-scale distributed searching has proven difficult in an Internet environment for large numbers of nodes (e.g., > 100).

The OAI retreats from the model of distributed searching, and attempts far less technical specification than previous DL interoperability projects. As a result of this decreased

scope, the OAI is proving to be a more flexible and resilient for interoperability - a sort of "RISC" (reduced instruction set computer) model for DL interoperability. The OAI defines only a generic bulk metadata transport protocol, and leaves other features to be borrowed from other technologies or implemented as independent services.

A special workshop, "Experimental OAI-Based Digital Libraries", was held at the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001), September 4-9, 2001, Darmstadt, Germany (www.ecdl2001.org). The purpose of this workshop was to bring together practioners and developers interested in building interoperable digital libraries based on the OAI protocol and principles. The twenty-nine workshop participants (listed in Table 1) came from ten different countries to hear a program of five invited presentations and five contributed papers.

Laura Maria Anselmi	<laura.anselmi@biblio.polimi.it>
Donna Bergmark	<bergmark@cs.cornell.edu>
Leona Carpenter	<l.carpenter@ukoln.ac.uk>
Leslie Carr	<lac@ecs.soton.ac.uk>
Roel de Cock	<roel@dtv.dk>
Avril Conacher	<avril.conacher@ed.ac.uk>
Susanne Dobratz	<dobratz@rz.hu-berlin.de>
Colm Doyle	<colm.doyle@lub.lu.se>
Christian-Emil Ore	<c.e.s.ore@muspro.uio.no>
Gudrun Fischer	<fischer@lsb.cs.uni-dortmund.de>
Sarah Fredline	<s.fredline@qut.eud.au>
Jochen Hollmann	<joho@ce.chalmers.se>
Louis Houle	<louis.houle@gouv.qc.ca>
Rick Luce	<Rick.Luce@lanl.gov>
John MacColl	<john.maccoll@ed.ac.uk>
Hamid Reza Mehrabi	<hrm@kb.dk>
Michael Nelson	<M.L.Nelson@larc.nasa.gov>
Irene Onyancha	<Irene.Onyancha@fao.org>
Anna Ortigari	<ortigari@cib.unibo.it>
Corrado Pettenati	<Corrado.Pettenati@cern.ch>
Adrian Price	<apr@kb.dk>
Michele Rubini	<Michele.Rubini@biblio.polimi.it>
Diann Rusch-Feja	<ruschfeja@mpib-berlin.mpg.de>
Gauri Salokhe	<schwander@lanl.gov>
Thorsten Schwander	<schwander@lanl.gov>
David Smith	<dasmith@perseus.tufts.edu>
George Spencer	<gaspence@indiana.edu>
Stefan Winkler	<stef.winkler@gmx.de>
Mohammad Zubair	<zubair@cs.odu.edu>

Table 1. Workshop Participants

Contents

Welcome & Introduction	5
<i>Diann Rusch-Feja, Max Planck Institute for Human Development</i>	
Knotting Together Digital Library Services: Progress and Challenges Ahead (Invited Talk)	10
<i>Richard Luce, Los Alamos National Laboratory</i>	
Convincing the Institution: Developing an Institutional Open Archive of Research Publications at the University of Edinburgh	27
<i>John MacColl & Avril Conacher, University of Edinburgh</i>	
Integrated Data Harvesting in the Perseus Digital Library System	34
<i>David A. Smith & Anne Mahoney, Tufts University</i>	
Heterogeneity in Open Archives Metadata	43
<i>Gudrun Fischer & Norbert Fuhr, University of Dortmund</i>	
Enhancing OAI Metadata for Eprint Services: Two Proposals.....	47
<i>Tim Brody, Zhuoan Jiao, Steve Hitchcock, Les Carr & Stevan Harnad, University of Southampton</i>	
Integration of Grey Literature with Electronic Journals (Invited Talk)	53
<i>Corrado Pettenati, CERN</i>	
The Use of Open Archives: Who, How Often and Why (Invited Talk)	71
<i>Leslie Carr, University of Southampton</i>	
Metadata Framework for Resource Discovery of Agricultural Resources	91
<i>Irene Onyancha, Fynvola Le Hunte Ward, Frehiwot Fisseha, Kafkas Caprazli, Stefano Anibaldi & Keizer Johannes, Food and Agriculture Organization (FAO) of the United Nations Library & Documentation Systems Division</i>	
OAI Experiences with Arc and Kepler	99
<i>Mohammad Zubair, Old Dominion University</i>	
European Support for Open Archive Activity: the EU-OAF Project (Invited Talk)	125
<i>Susanne Dobratz, Humboldt University</i>	
Closing Remarks	133
<i>Diann Rusch-Feja, Max Planck Institute for Human Development</i>	

*European Conference on Digital Libraries (ECDL)
Darmstadt, 8 September 2001*

***Experimental OAI Based Digital
Library Systems -
Workshop for OAI Implementers***

Diann Rusch-Feja

**Library and Research Information
MPI for Human Development, Berlin**

*Experimental OAI Based Digital Library Systems
(ECDL Workshop)
Darmstadt, 8 September 2001*

Workshop Organizers

**Kurt Maly (ODU)
Mohammed Zubair (ODU),
Michael Nelson (NASA)
Diann Rusch-Feja (MPIB)**

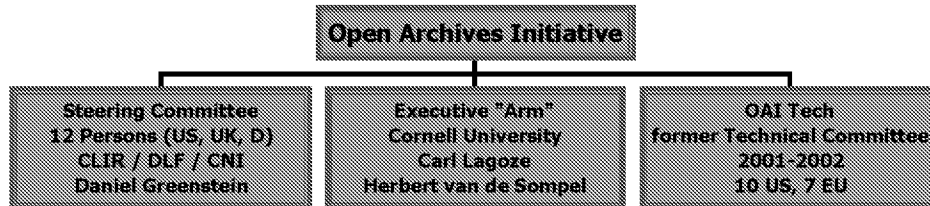
Open Archives Initiative

- Heighten Visibility of Research and Publications
- Connect Scholarly Pre/E-print Servers, Navigation Tools (Interoperability) and Enable New Services
- Contribute to Enhanced Information Access for Researchers, Faculty, Scholars, Students
- Respond to Changes in the Research Process, Publication Technology, Scholarly Communication
- Integral Component of the Digital Library Concept and of Research in DL, IR and Scholarly Communication

Open Archives Initiative

- [Http://www.openarchives.org](http://www.openarchives.org)
- Workshops 2000 ACM / ECDL
- January 2001 - OAI Metadata Harvesting Protokoll Ver. 1.0 published
- January / February 2001 - OAI Information Days (Washington, D.C., / Berlin, Germany)
- Experimental Phase 2001-2002
- CERN OAI Workshop 22-24 March 2001 in Geneva - OAI and Open Peer Review

OAI Organigram



Organisation der OAI

- Supported by CLIR, Digital Library Federation (DFL), CNI, NSF (USA)
- Steering Committee (12 members, 2 EU)
- Technical Committee (ad hoc in 2000), 2001 Expansion for ca. 1 year - OAI-Tech
- OAI Executive - Cornell University, USA - Carl Lagoze, Herbert van de Sompel
- Europa: DINI and DFG (D), JISC (UK), EU

Division of Responsibilities

Steering Committee

- Political & Strategic Decisions
- Promotion & P.R. etc.

Executive

- Coordination of FAQs, Website, Mailinglists
- Registry & Implement.
- Coordination of TC
- Organize Events

Technical Committee

- Protocol Testing u. Modification (international)
- Project Assistance
- Registry Services, Implementers
- Metadata Standards
- Support f. OAI Executive
- Support for individual Subject Communities

Reasons for Expanding the OAI-Tech 2001

- Need for establishing a core group with technical OAI expertise but subject/community differentiation
- Focus direct interaction between OAI implementers and developers
- Feedback and Integration of the OAI Alpha Users for improving the Specifications
- Need for greater international participation and direct input in OAI Development & Support

Goals of the Workshop

- Provide Interaction with Implementers
- Discuss and Identify Community / Regional Issues
- Investigate and Define Open Issues to Take to the Technical Committee
- Expand the OAI Network –
 - Information
 - Promoting Adoption and Informal Support
 - Stimulating New Ideas / Areas of Application & Development

Knotting Together Digital Library Services:

Progress and Challenges Ahead

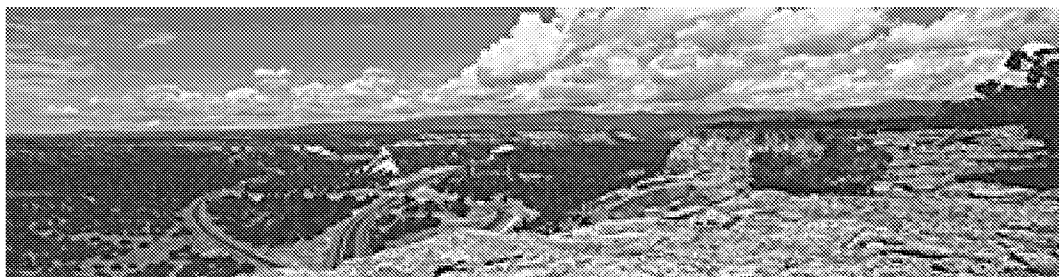
Rick Luce
Research Library Director
Library Without Walls Project Leader
Los Alamos National Laboratory

ECDL 2001 Workshop: Experimental
OAI Based Digital Library Systems
September 8, 2001



Outline

- ◆ Contextual background
- ◆ Our experience today with OAI
- ◆ Challenges ahead
- ◆ What comes next ??



Research Library

Mission: *Increase research productivity*

- ◆ Hybrid - academic science & special library
- ◆ 8,600 active Los Alamos customers
 - ✓ 200,000 external users in 29 institutions
- ◆ Staff = 52 (includes 10 on *Library Without Walls* team)

Vision: *To create a network of knowledge systems that facilitate scientific communication and collaboration*



**1999 Federal Information
Center of the Year award**

Los Alamos Research Library 9/01

User Needs

Assumptions: Given research is a race against time....

- ✓ Scientists need convenience, accessibility, and high quality content from a single interface
- ✓ The library mission must support increasing the efficiency of science & technology research

Los Alamos Research Library 9/01

Challenge: On a Large Scale...

Provide access to diverse, cross-discipline e-print collections

- ◆ **Interoperability:** locating relevant content among heterogeneous & variable systems
 - ✓ *Use these systems as one virtual collection*

Los Alamos Research Library 901

Open Archives Initiative: Premise

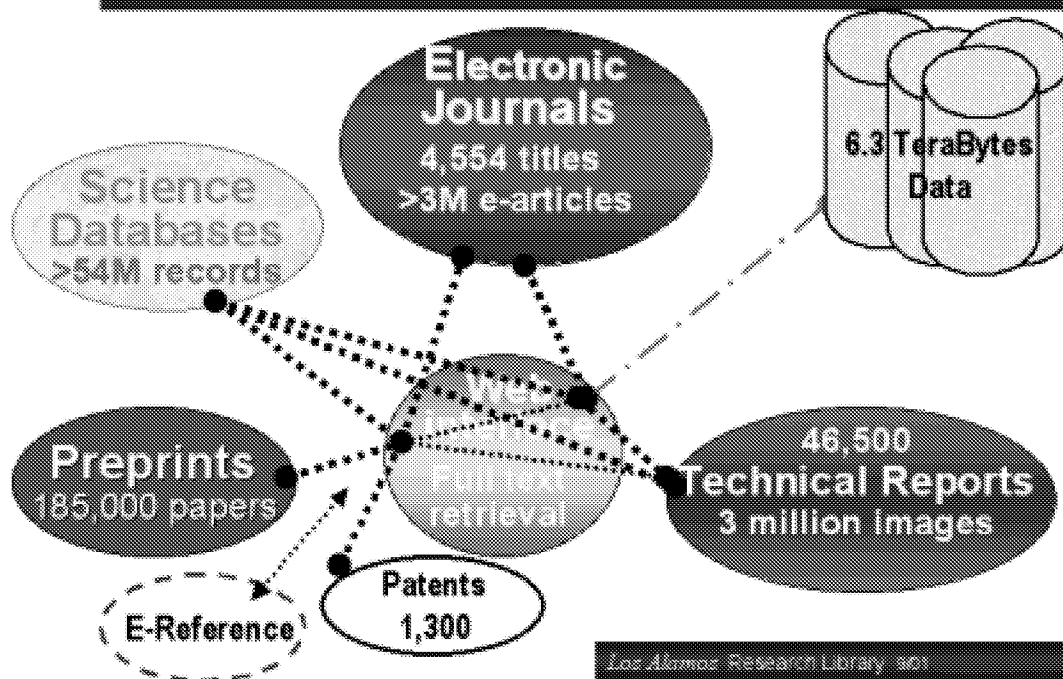
Goal: *Catalyze progress in new scholarly publishing models over next 5-10 years*

- ◆ Create a universal service for non-peer reviewed scholarly literature*
 - ✓ Fundamental and free layer of scholarly communication
 - ✓ On top lies free and commercial services

* Summer 1999: Ginsparg, Van de Sompel, Luce

Los Alamos Research Library 901

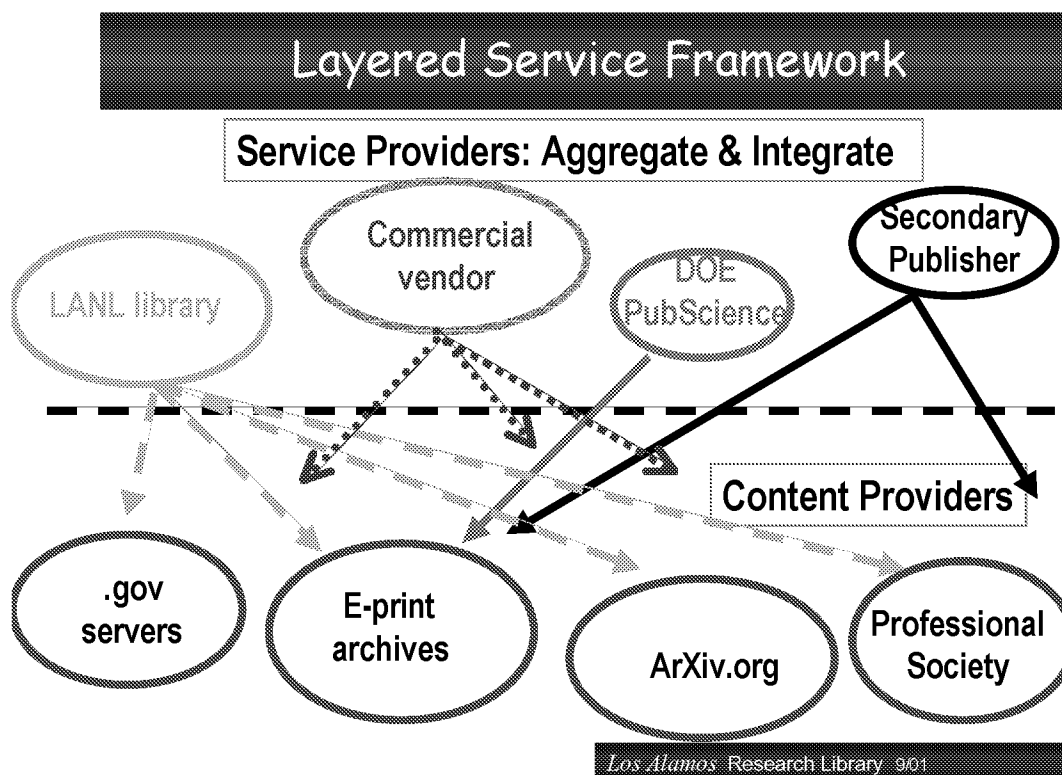
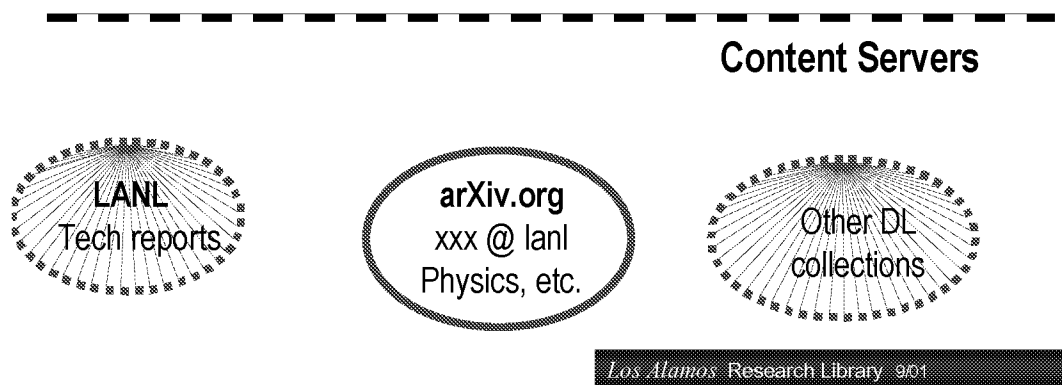
Library Without Walls - Local Digital Content



OAI Experience Today

- ◆ arXiv.org – OAI compliant
- ◆ TRI (Technical Report Interchange) project with ODU
 - ✓ NASA, LANL, AFRL tech reports
- ◆ Waiting funding for:
 - ✓ OAI-Compliant Federated Physics Digital Library for the NSDL
 - Collaboration with ODU and APS

Open Archives Framework



Adding Value

- ◆ Enhancing capabilities
 - Discovery tools on heterogeneous collections
 - Rich, dynamic linking
 - Personal alerting
 - Reviews and notation
 - Citation analysis
 - Recommendation systems

Los Alamos Research Library 901

Integration : Multi-database search

The screenshot shows the FlashPoint search interface. At the top left is the Los Alamos Research Library logo with a 'Library home' link. The title 'FlashPoint' is in a large, stylized font, with the subtitle 'A multi-database search tool v.1.1.1' below it. The 'Quick Search' section features a search input field with the text 'optics' and a 'Search' button. Below the input field are radio buttons for search criteria: 'Author' (selected), 'Title/Abstract/Keyword', and 'Source/Journal'. A note below the radio buttons says '[Author - enter Last name First initial]'. Underneath are nine checkboxes for different databases: 'AI' (checked), 'BIOSIS', 'DOE Energy', 'Engineering Index', 'e-Print arXiv', 'INSPEC', 'Science Server (B-journals)', 'SciSearch', and 'Social SciSearch'. A 'Database' button is located below the checkboxes. At the bottom, the Los Alamos National Laboratory logo is on the left, and the text 'Operated by the University of California for the US Department of Energy' is in the center. A footer bar at the bottom right contains the text 'Los Alamos Research Library 901'.

Los Alamos Research Library

Library home

FlashPoint
A multi-database search tool v.1.1.1

Quick Search

? Enter your search:

(1998 - Present) ☐ Author ☒ Title/Abstract/Keyword ☐ Source/Journal
[Author - enter Last name First initial]

☒ AI ☐ Engineering Index ☐ Science Server (B-journals)
☐ BIOSIS ☐ e-Print arXiv ☐ SciSearch
☐ DOE Energy ☐ INSPEC ☐ Social SciSearch

Los Alamos
NATIONAL LABORATORY


Operated by the University of California for the US Department of Energy
(copyright © UC 2000) | Notice to Users/Disclaimer

Los Alamos Research Library 901

Multi-database search results



Research Library




FlashPoint
A multi-database search tool v1.1

Quick Search:

Database	Number of Results	
SciSearch @ LANL	3045	[new window]
SciSearch @ LANL	59731	[new window]
Science Server @ LANL	39023	[new window]
Science Search @ LANL	300	[new window]
INSPEC @ LANL	108719	[new window]
Engineering Index @ LANL	60131	[new window]
DOE Energy @ LANL	5045	[new window]
BQSIG @ LANL	7696	[new window]

Note: if your results do not show all the databases you selected, please click [here](#) to reload this page

[New Search](#)
[Help](#)
[Comments](#)




Los Alamos
NATIONAL LABORATORY


Operated by the University of California for the US Department of Energy
Contract # DE-AC02-89OR21400 | Network to Users' Materials

Los Alamos Research Library 901


Direct connection to relevant database



Los Alamos Research Library



DOE Energy
at LANL



FlashPoint
A multi-database search tool v1.1

DOE Energy Retrieval Results

5045 out of 142973 records matched the query below. 50 records displayed
optics in title and year >= 1998 and year <= 2000

Marks	Score	Title, Author, Source
<input type="checkbox"/>	1.00	Apparatus and method for the determination of group velocity in thin film Hsieh, H. I. Patent No. US 6,016,016/99; Priority No. US patent application 8,110,885; Assignee: Brown Univ. Research Foundation, Providence, RI/029; Date Filed: 7 Jul 1999
<input type="checkbox"/>	1.00	Layer structure and novel system for data processing Hsieh, H. I.; Hwang, J.; Kuo, H. H.; Chen, T. B. Patent No. US 6,017,103/99; Priority No. US patent application 8,844,772; Assignee: Univ. of California, Oakland, CA/035; Date Filed: 17 Jan 1999
<input type="checkbox"/>	1.00	Polymers used in thin film for non-linear optical materials Yang, X.; Li, D. Q.; Swenson, B. I. Patent No. US 6,017,773/99; Priority No. US patent application 8,844,772; Assignee: Univ. of California, Los Angeles, CA/035; Date Filed: 18 Apr 1999
<input type="checkbox"/>	1.00	Optical waveguide for data processing Tobias, D. A.; Swenson, B. I.; Hsieh, H. I.; Kuo, H. H. Patent No. US 6,017,284/99; Priority No. US patent application 8,110,885; Assignee: Brown Univ. Research Foundation, Providence, RI/029; Date Filed: 23 Sep 1999


Quick Search:

Number of Results

3045	[new window]
59731	[new window]
39023	[new window]
300	[new window]
108719	[new window]
60131	[new window]
5045	[new window]
7696	[new window]

Note: if your results do not show all the databases you selected, please click [here](#) to reload this page

[Comments](#)



Los Alamos
NATIONAL LABORATORY

Operated by the University of California for the US Department of Energy
Contract # DE-AC02-89OR21400 | Network to Users' Materials

Los Alamos Research Library 901

BIOSIS Record

Los Alamos Research Library xml.lanl.gov

[Up](#) [Prev Rec](#) [Next Rec](#) [New Search](#) [Help](#)

Title: Biology back issues free as publishers walk highwire. [[Bio](#)]
Author: Butler, Dorian [[Bio](#)]
Journal: Nature (London), March 9 2000, v.404, no.6774, p.117. [[Bio](#)]
Journal Abbr.: NATURE (LOND) [[Bio](#)]
Abstract: High Wire Press, a not for profit organization, will collaborate with 83 life science journals to provide free access to back issue articles on the World Wide Web. [[Bio](#)]
Concept: Information Studies [[Bio](#)]
Organization: High Wire Press -- company/organization, not for profit organization [[Bio](#)]
Miscellaneous: back issue articles [[Bio](#)]; life science journals -- free online access [[Bio](#)]; News Article [[Bio](#)]; World Subject: Wide Web [[Bio](#)]
Doc. Type: News [[Bio](#)]
Language: English [[Bio](#)]
ISSN: 0028-0836 [[Bio](#)]
CODEN: NATUAS [[Bio](#)]
Control No.: PREV200000243565 [[Bio](#)]
SICI: 0028-0836(2000)404:6774<117BKIFAP>2.0.TX;2- [[Bio](#)]

Contains material from the following databases:

(Click name to list record display to a single database)

[RTVCPC](#)

Los Alamos Research Library 901

SciSearch record

Los Alamos Research Library xml.lanl.gov

[Up](#) [Prev Rec](#) [Next Rec](#) [New Search](#) [Help](#)

Title: Biology back issues free as publishers walk HighWire [[Sci](#)]
Author: Butler, D [[Sci](#)]
Journal: NATURE, MAR 9 2000, v.404, no.6774, p.117-117 [[Sci](#)]
Journal Abbr.: Nature [[Sci](#)]
Publisher: MACMILLAN MAGAZINES LTD [[Sci](#)]
Publ. Address: PORTERS SOUTH, 4 CRINAN ST, LONDON N1 9XW, ENGLAND [[Sci](#)]
Publ. Date: MAR 9 [[Sci](#)]
Journal Cat.: D RO MULTIDISCIPLINARY SCIENCES [[Sci](#)]
Doc. Type: Miscellaneous [[Sci](#)]; News Item [[Sci](#)]
Language: English [[Sci](#)]
ISSN: 0028-0836 [[Sci](#)]
CODEN: NATUAS [[Sci](#)]
Control No.: A2000293UC00015 [[Sci](#)]
SICI: 0028-0836(20000309)404:6774<117BBIFAP>2.0.TX;2- [[Sci](#)]

Contains material from the following databases:

(Click name to list record display to a single database)

[SciSearch](#)

Special E-Books by Ex Libris (USA) Inc. - Netscape

LinkOeeker **LANL Research Library**

Title: Identification of a nasal gene on chromosome 7q32 that is interrupted by a translocation breakpoint in an autistic individual

Source: AMERICAN JOURNAL OF HUMAN GENETICS [0002-9297] year 2000 vol: 67 iss: 2 pp: 510

Press Ctrl-F to prefill this screen.

① Full-text article
year: 2000 volume: 67 issue: 2 start page: 510

② Request this Document

③ Find other articles by this author in FlashPoint
Select author: Vincent, John B

④ Is this paper cited?
☒ SciSearch @ LANL ☐ Social SciSearch @ LANL

⑤ Is this author cited?
Select author: Vincent, John B
☒ SciSearch @ LANL ☐ Social SciSearch @ LANL

⑥ Get sequence information from PubMed?
name entry: AC017647

Document Download

BIOSIS at LANL

Search Results

below 50 records displayed
<?and>year <= 2001

... variants from SHIRAH and YOKAWA and the TT gene-like

... genyil, Mishra, Shunji
99-123

... is interrupted by a translocation breakpoint in an autistic

... V, Botten, Patrick F., Roberts, Wendy, Scherer, Stephen W.
2000; v.67, no.2, p.510-514

... gene is a bona fide secretion pathway dependent

... nuclear localized double-stranded DNA binding protein.
Yang, Bing, Zhu, Weiguang, Johnson, Lowell B, White, Frank F.
Source: Proceedings of the National Academy of Sciences of the United States of America, August 15 2000, v.97,
no.17, p.9807-9812

Save this Article

Google My Yahoo! Bookmarks Fickrhome AddToMylb++ MyLb++

Navigation Contents Home

Issues in Science and Technology Librarianship Winter 2001

Indexed article

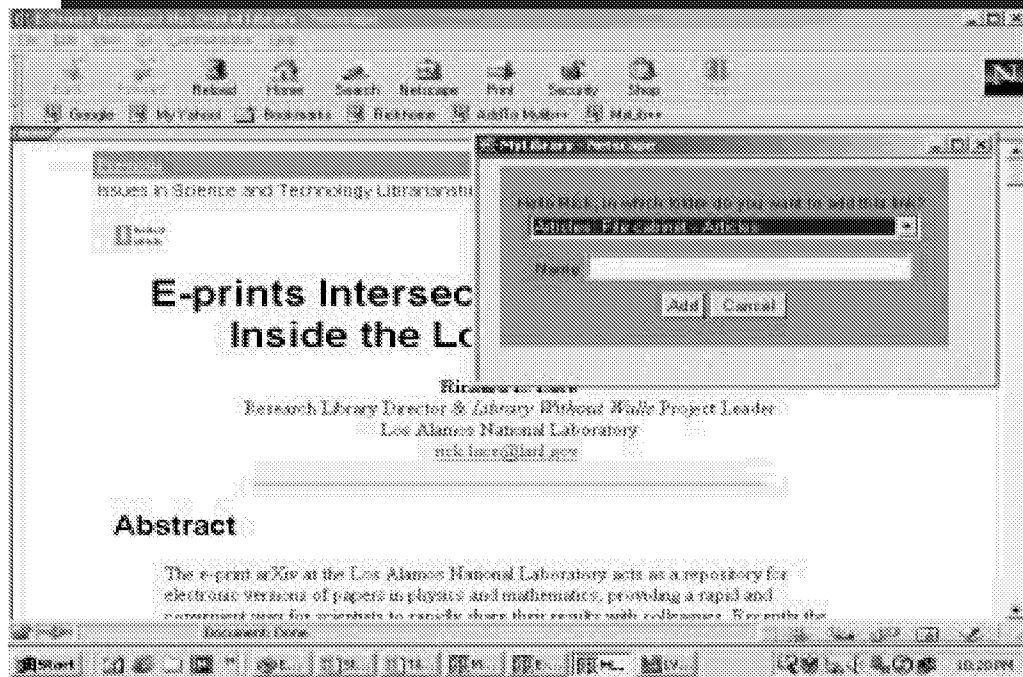
E-prints Intersect the Digital Library: Inside the Los Alamos arXiv

Richard E. Luce
Research Library Director & Library Without Walls Project Leader
Los Alamos National Laboratory
rick.luce@lanl.gov

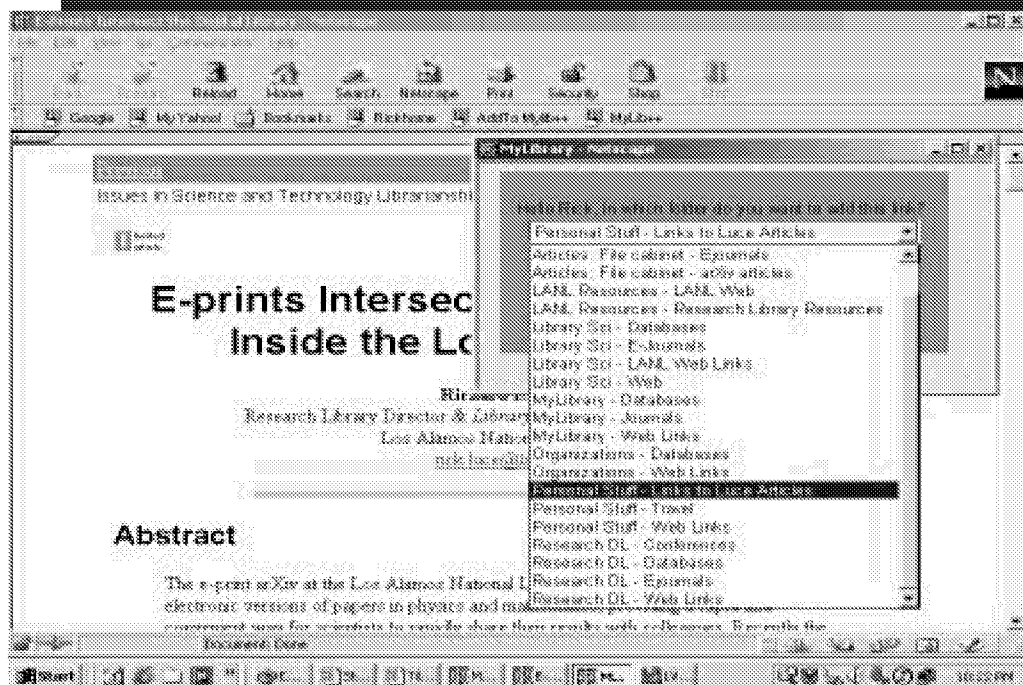
Abstract

The e-print arXiv at the Los Alamos National Laboratory acts as a repository for electronic versions of papers in physics and mathematics, providing a rapid and convenient way for scientists to readily share their results with colleagues. Recently the

Using Bookmarklets from the Browser



Choose a File Folder and Name



Create Shared User Group in MyLibrary

MyLibrary - LANL Research Library - Netscape

File Edit View Go Communicator Help

Los Alamos National Laboratory

Lab Home | Phone | Search

RESEARCH LIBRARY

Site Map MyLibrary Catalog Contact Us Services Submit Requests What's New

Add Library

| | | |
|----------------------------------|---|--|
| Title | <input type="text" value="TRI Project"/> | Library title. Will be shown in the Tab at the top. |
| Type | <input type="radio"/> Private <input checked="" type="radio"/> Shared | Library Type. Defines a private or shared library. |
| Shared Library Read-Only Users | <input type="text" value="Doug, Mark, Miriam, Ming"/> | Users (members) belonging to a shared library. Please, provide their user names, separated by space. |
| Shared Library All-Right's Users | <input type="text" value="Francis, L&I"/> | Users (members) belonging to a shared library. Please, provide their user names, separated by space. |
| Area of Interest | <input type="text" value="Grants & Funding"/> | Messages and default folders are based on this selection. |

Los Alamos Research Library 901

Shared Folder in MyLibrary

MyLibrary - LANL Research Library - Netscape

File Edit View Go Communicator Help

Los Alamos National Laboratory

Lab Home | Phone | Search

RESEARCH LIBRARY

Site Map MyLibrary Catalog Contact Us Services Submit Requests What's New

Welcome Bob

MyLibrary

LANL

[Add Library] [Advanced Features] [Sign out]

TRI Project MyLibrary LANL Resources Library Sci Research DL Personal Stuff Interesting Links File cabinet

Databases

COS funding opportunities (Funding opportunities)

CRISP

CEERIX

Default links for Grants

[Help?] [Comments]

Los Alamos Research Library 901

Next: Add Kepler to MyLibrary



Personalization

A mechanism to enable communication between users, agents, and information resources leading to information exchange, adaptation and recombination

1. **Requires unique user identity**
 - ✓ Authentication
2. **Knowledge of user behavior**
 - ✓ Personal preferences and usage statistics
3. **Knowledge of *communities of interest***
 - ✓ Behavior of relevant fields or communities

Scientific Communication

We need a mechanism to enable communication between users, agents, and information resources leading to information exchange, adaptation and recombination

Los Alamos Research Library 901

Active Recommendation Systems

Self-organizing knowledge on distributed networks driven by human interaction

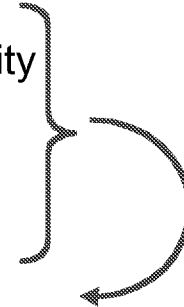
- ♦ A means to recognize users (agents)
- ♦ A means to characterize information resources
- ♦ Conversation mechanisms between users and information resources
- ♦ Adaptation mechanisms

Los Alamos Research Library 901

Adaptation of Structure and Semantics - Using Collective Behavior of Users

1. Knowledge contexts categorized
 - ✓ Keywords & keyword semantic proximity
 - ✓ Citations and citation proximity
 - ✓ Semantic proximity
 - ✓ Traversal proximity
2. Recommendation(s) calculated
3. Traversal proximity analyzed
4. Adaptation in system

Users + Profiles = learning community



Los Alamos Research Library 901

Recommendation Systems with Shared Knowledge Models

From Experts

List of Seminal
Articles

"PRECONDITIONED
ITERATIVE SOLVERS
FOR LINEAR SYSTEMS"



Expansion via Citation
Structure



Shared Knowledge
Structures

Graphs of associations among
keywords or among documents
from Bolten's user model
methodology.

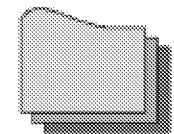


Los Alamos Research Library 901

Recommendation Systems with Shared Knowledge Models

Characterization of User Communities

From User Logs



User Logs
Documents accessed by users in a period of time (month)



Inferred Communities
Journal Titles that tend to be selected by the same users

Shared Knowledge

Structures
Graphs of frequency associations among keywords or among documents

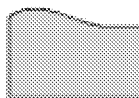
Users/Agents
Typical path behavior on SKS observed in Journal Clusters

Compile graph characteristics of SKS for simulations

From Experts

List of Seminal Articles

"PRECONDITIONED ITERATIVE SOLVERS FOR LINEAR SYSTEMS"

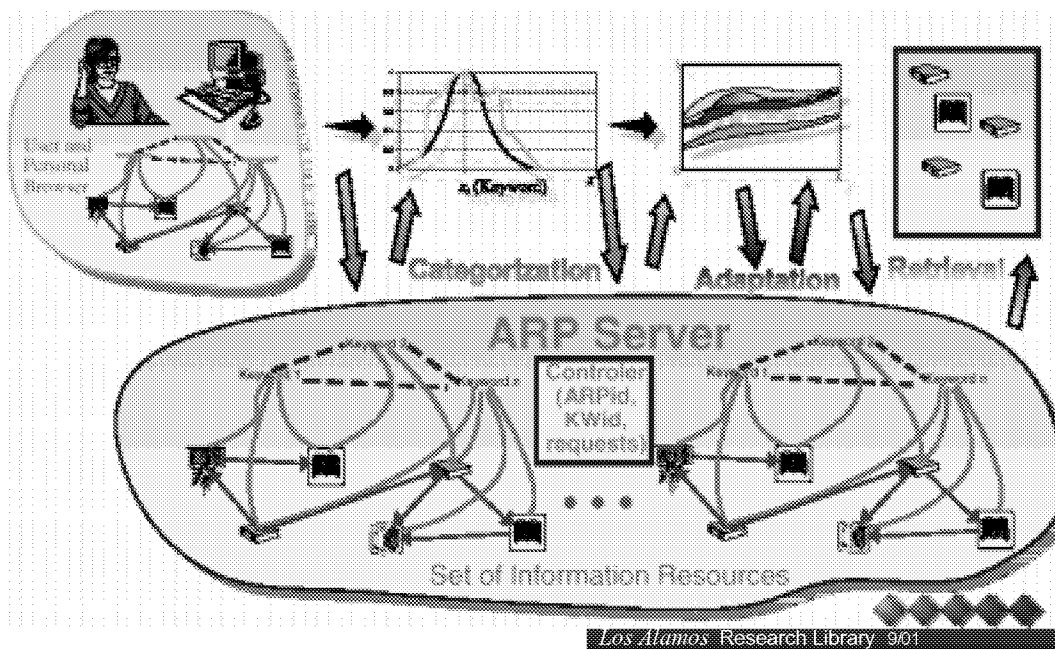


Expansion via Citation Structure

Shared Knowledge Structures
Graphs of associations among keywords or among documents from Bollen's user model methodology

Los Alamos Research Library 901

LANL Active Recommendation System



Conclusions: What Have We Learned ?

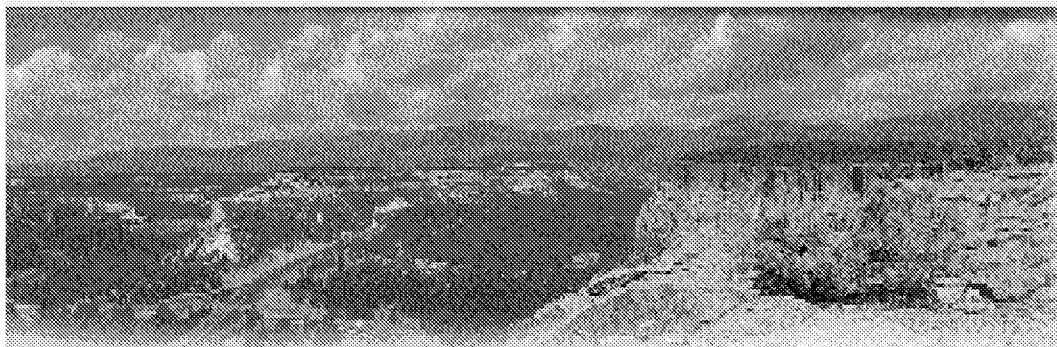
1. Digital library development is not just a technology challenge
 - ✓ Institutional structures require hybrid organizations capable of rapid re-adaptation
2. Requires Information Science in the broadest sense
 - ✓ IT, computational & computer sciences, library science, psychology, sociology, cognitive science, human factors ...
3. Organizations must manage information as a basic utility
 - ✓ CIO approach to organizing knowledge assets
 - ✓ Requires a unified information architecture
4. Foster the creation of institutional & society author repositories (options are not mutually exclusive)
5. Augmentation through external collaboration

Los Alamos Research Library 9/01

Discussion -- Questions ??

rick.luce@lanl.gov

<http://lib-www.lanl.gov/>



Convincing the Institution: Developing an Institutional Open Archive of Research Publications at the University of Edinburgh

John MacColl and Avril Comacher
*SELLIC (Science & Engineering Library,
Learning & Information Centre) Project*
University of Edinburgh
Darwin Library
The King's Buildings
Mayfield Road
Edinburgh EH9 3JU
john.maccoll@ed.ac.uk
avril.comacher@ed.ac.uk

Introduction

This paper examines the process of creating an archive of research publications within a single, large research-oriented university. At the University of Edinburgh, we are in the process of creating such an archive at the present time, and are currently working with the Faculty of Science & Engineering, which is the largest faculty in the University, with approximately 7,000 students. Other faculties will be included in the archive at a later date, assuming additional funding can be made available. The creation of the archive is being led by the library, which is fully committed to the idea of an open research archive sustained by self-archiving. Creating a shared vision of this is difficult in several ways.

Background

The University of Edinburgh is an ancient Scottish university, the largest in Scotland and one of the so-called 'Russell Group' of research-led universities within the UK. It presents very ripe territory for a self-archiving culture based on the Open Archive Initiative. The university produces a substantial quantity of research. Recent work within the Library in supporting the institution's submission to the current UK Research Assessment Exercise (RAE) resulted in an estimate that the total number of research outputs by university academic and research staff is around 3,750 items each year.

OAI and Self-Archiving

Within the Library at Edinburgh, our view is that the practice of self-archiving of research publications is vital to the efficiency of scholarly communication in what Stevan Harnad calls the 'post-Gutenberg' age. The assertion by academic authors in the university of the right to mount copies of their own research publications on a university server which is open to fellow-researchers across the world, if extended thoroughly into all disciplines, would mean that research output was available to peers in the shortest possible timescale – an objective long-sought by most disciplines. From the library's perspective, it would also and very importantly allow university libraries to cancel journal subscriptions, as it became clear that the titles available by subscription – whether in print or electronic formats – were being used less and less because their contents were already available on university servers where they could be easily accessed.

What happens beyond that point is not clear. If the University of Edinburgh were to cancel even one-third of its journal titles, for example, it would realise an annual saving of over £100,000 – the greater part of which we would be expected to return to the university's central coffers. However, the possibility of academic journal publishers going to the wall if

the pattern of cancellations at Edinburgh were replicated in a large number of academic libraries is not one which libraries would really like to see, and academics certainly would not. Somehow, then, we require to keep journal publishers in business as providers of quality control services – whether still on the basis of established journal titles, or by some other means. That implies the use of some part of the £100,000 saving in order to pay for quality control – potentially both for articles which go on to be published, and for those which don't. At the present time, most publishers are not genuinely fearful of losing subscriptions on a massive scale – though many are offering new products based on digital content, which may be an attempt to diversify ahead of the cancellation of core products.

A system of scholarly communication altered in this way would represent a 'win-win' outcome for scholars and their institutions (via their libraries). The key to getting there is author self-archiving. This logic seems relatively straightforward to us, but of course establishing a system of self-archiving of research publications requires academics to change working and publishing habits, and it is when one begins to address these points to the academics in departments and faculties that it becomes clear that the case for self-archiving, so obvious to us, is not so obvious to them. One reason is that the savings benefit to the institution does not translate directly into increased funding for them. The indirect benefit, to their share of the library budget, should be becoming more and more visible to them as each successive year pushes journal prices up by an average 9% or 10%, thereby reducing the purchasing power of their budget shares. Nevertheless, for many researching academics able to find supplementary cash to shore up the subscriptions for the journals essential to their work, the loss of funds for textbooks for teaching purposes is not a very real problem.

Then there are other concerns. The familiarity with preprints, enjoyed by physicists and mathematicians, is not necessarily shared by other disciplines. Many biological scientists, in particular, are uncomfortable with the notion that pre-published articles should be freely available on the web. Some have voiced fears about their work 'being scooped' by competitors in other institutions, and others have described the 'publication lag' – the period between an article being accepted for publication and its eventual appearance in print – as positively valuable, since it presumably allows some early commercialisation of product before the research hits the public domain. Comments of this sort are interesting, and illustrate how varied the scholarly publishing world can be, but in the end only obfuscate the issue. What self-archiving seeks to do is to make more efficient the process of scholarly communication. Whenever the academic wishes that communication to occur, the self-archiving model is the best means of achieving it. Using the inefficiencies of print publishing as a means of exploiting research priority is purely opportunistic. In a more efficient system, other means of delaying publication, if that is what an author wishes to do, would have to be found.

Self-archiving, of course, need not be concerned only or even mainly with preprints. The distribution of preprints is as much an option for researchers in the digital age as it ever was in the print age. Nevertheless, our experience is that, in discussing self-archiving and its various components of preprints and postprints, copyright assertion and free availability of publications within a distributed, virtual database, many academics find the picture confusing, and become preoccupied with a particular component – such as preprints – and anxious about any implications for changing their own publishing habits. For librarians, it is important that we recognise how powerful are the levers within the academic reward culture. Professional recognition, as well as promotion, depend upon being published in recognised journals, and even though self-archiving need be no more than simply an adjunct to existing practice, many immediately perceive an attempt to substitute a tried and tested culture with something new and untested.

This is one of several reasons why we believe libraries should be involved in the machinery of self-archiving. As trusted administrators of scholarly resources, libraries can in effect take

over the administration of self-archiving, leaving academics with a scholarly publishing system which looks virtually identical to the one they have known all of their research careers. What we have found, therefore, in our attempts to establish an eprint archive at Edinburgh over the past few months, is that particular strategies are required to convince the institution that self-archiving of research publications is beneficial to them. The keys to making progress in introducing self-archiving in UK universities are listed in the next section.

Strategic Objectives in Securing Support for Self-Archiving

First, the support of senior figures in the university administration – including senior academics such as faculty deans and the convenor of the Research Committee – must be secured. Linking self-archiving to the RAE is clearly very sensible in this endeavour. Similarly, it is worth pointing to the fact that DAI presents opportunities for universities to make publicly available a range of different types of institutional publications, effectively providing a ‘shop window’ facility. While research publication must be the priority, administrators can be positively influenced by the idea that the same architecture can be used to host a range of different types of university publication on the web.

Second, we must respect the heterogeneity which exists within different discipline communities. Physicists are different from biologists. Historians and literary scholars are different from each other and from divines and lawyers. Librarians must be careful to avoid sweeping assumptions.

There is a need to emphasise the central resources which the library can call upon to take work away from academics. Freeing more time for research is a winning strategy.

Then we must stress the conservatism in the proposal. With self-archiving, the reality for most academics will be that their scholarly communication practices will scarcely be changed, and they can certainly continue to submit articles for publication to the top peer-reviewed journals in their field. Alongside this, it may also be necessary to emphasise that self-archiving will not be accompanied by a simultaneous exercise in cancelling existing journal subscriptions. The library will cancel on the basis of changes in the reading habits of scholars, not in order to force such change.

We must focus on peer-reviewed research articles for publication. In other words, it is important at this stage to keep academics on familiar territory. A willingness to deposit preprints is a bonus.

Finally, we have to acknowledge the importance of key journals, and the positive feelings many academics have towards the top journals in their disciplines. Academics do not look first and foremost at the price tags of journals, as librarians tend to do. Some have tremendous loyalty to particular journals, and it is important then to point out that all we may wish to change in the way the university deals with such a journal is the terms of copyright transfer.

The Research Assessment Exercise

Any self-archiving project concerned with research publication in the UK has to be set within the context of the RAE. Whether for good or ill, the RAE has come to dominate the culture of UK university research publishing. It has been in use now for around two decades, and has been modified in the course of that time, from an initial emphasis on the quantity of research published by academics, to the current emphasis upon the quality of research output. Universities submit their ‘research-active’ staff, each of whom puts forward their four most significant research publications (or other research outputs, such as performances in the case of academics in performing arts subjects, for example) over the four-year period of assessment. National assessment panels then consider the strength of these publications and

grade university departments accordingly. The exercise represents a huge administrative burden on universities in the UK – particularly on large ones like Edinburgh. There is first of all the business of deciding which staff should be described as ‘research-active’, and this involves some nice decisions. Academics who produce research which may not be considered of high enough quality for a department’s aspirations can be left out of the process. This decision can lead to humiliation and resentment on the part of these staff, but is nevertheless a sensible one in the context of an assessment logic which is formula-based, and grades departments by the quality of their collective research and the size of their submission. Departments are penalised for submitting a large number of researchers whose overall submission quality may be brought down by a few ‘sub-standard’ researchers. At the same time, they cannot resort to a strategy simply of submitting a few of their best researchers for assessment, since size of submission is taken into account.

In the most recent RAE submission, the University of Edinburgh submitted some 6,000 research publications, representing four publications from each of 1,500 research-active members of staff. The library had already established itself as having an administrative role to play, making available shelf space for the 6,000 submitted items, and creating a catalogue database for them, to support the work of faculties in dealing with the assessment panels. It proved very effective, and an open archive approach next time round, allowing not just the metadata but the full-text of the article itself to be retrieved on demand by assessors, is enormously attractive to the administrators on the ground.

Institutional or Disciplinary Archives?

Those academics who have experience of eprint archives tend only to consider the discipline-based archive approach. The largest eprint archive in the world, and the originator of many of the attributes of the self-archiving movement, is arXiv.org, which until recently was based at the Los Alamos National Laboratory. Many researchers in physics and cognate disciplines are users of arXiv. Since researchers in general belong first to a discipline and then to an institution, the disciplinary archive feels more natural to them than does an institutional archive. Convincing them of the value of the institutional archive approach involves making the point, first of all, that the two types of archive are not mutually exclusive. Research publications can be archived both in a disciplinary and an institutional archive. The institutional archive is an insurance policy that guarantees that the publication will be available somewhere within the Open Archive network. Secondly, an appeal can be made via those disciplines which as yet have no or little tradition of using disciplinary archives. Rather than wait for one to appear, and then have no guarantee that it will not vanish off the web when its originator departs the scene, these academics can ensure that their work is archived now, with minimum effort, by their library.

One objection which has been made to the institutional approach is that, in the case of multi-institutional co-authored papers, it is not clear whose institution may lay claim to them. This is a real problem, albeit a minor one, but as a reason to prefer disciplinary archives over institutional archives it is surely spurious. In the final analysis, scholarship will be served as long as the eprint appears on somebody’s archive, somewhere. The linkage of the institutional archive methodology to the RAE and therefore to the general research profile of a university does however require that this question be sensibly addressed. The RAE methodology allows for discretion by assessors in the case of co-authored works, so there is no obvious definition to draw upon in determining the ‘lead’ researcher in a research collaboration. Unlike the RAE, however, where the restriction to four outputs per researcher would probably cause institutions to be somewhat wary of submitting an output on behalf of a researcher whose role may have been a minor one, in the case of institutional archives of research, the opposite may be true. Ultimately, if the same widely co-authored article is stored in six or seven different archives, this should not matter, and will be a problem for the OAI search service software products to handle. It is a manifestation of the problem researchers have long known in using

union catalogues, and is not difficult to resolve. One way of resolving it would be for individual institutions to use automated resolver tools, which are described in the next section.

The Role of the Library

Academic staff do not necessarily see the library as having a role to play in this area (indeed, nor do some librarians). Nevertheless, there is a certain logic to locating the responsibility for such an archive in the library. University libraries perform two particularly relevant tasks, which have been part of their mission throughout their history: bibliographic description and document preservation. The skills to describe publications authoritatively, and the emerging technology of digital preservation, therefore already exist within the library.

At a more strategic level, the library pays the bills for the print and electronic journals it purchases on behalf of the university. This allows it to monitor usage patterns of these journals, and the growth of self-archiving should in due course lead to a reduction in demand for the subscription-based titles, which the library will observe and which will influence its recommendations to departments regarding cancellation of titles.

A recent development in the architecture of digital libraries is the use of 'resolver services', which resolve the links followed by users as they navigate through digital documents to publications which the library can provide either in full-text, because it has a subscription or free access to the relevant service, or in surrogate form – i.e. to a catalogue record for a book or journal available in hard copy on a library shelf, or, indeed, to a request to the library's document delivery service. Resolvers will allow libraries to choose between two digital versions of the same article. As open research archives grow, libraries can select whether to resolve links for published articles to an OAI-compliant archive, or to a commercially-published service, if they prefer. This ability will provide a very powerful tool in the hands of librarians in directing the attention of students and researchers to the free research corpus. Resolver services may thereby accelerate the process of widespread institutional self-archiving – but they will require to be adopted by librarians whose motive will be mainly financial.

Researchers may not see it appropriate that the library should handle the copyright negotiation between them and publishers, and in this area, practice is likely to vary from institution to institution. One good reason for the library to handle this is because it allows it to intervene in the scholarly publishing process at the point at which publication is imminent, and so capture a research article for the local archive if the author has not already submitted it at that point. And indeed libraries do have experience in copyright negotiation. Recent years have seen the arrival of 'digital readings' (usually known, somewhat unhelpfully, as 'electronic reserves') within library services. Obtaining permission from publishers to digitise an article for hosting in such a service requires negotiating a fee with them. Our experience at Edinburgh has revealed that some departments already have policies governing copyright retention, and at least one uses a form of wording which is more assertive than most of us would dare to claim:

The «research sponsor» and the University of Edinburgh are authorised to reproduce and distribute reprints and on-line copies for their purposes notwithstanding any copyright annotation hereon.

In general we would adopt a more conciliatory form, along the lines advocated by Steven Harnad:

I hereby transfer to [publisher or journal] all rights to sell or lease the text (on-paper and on-line) of my paper [paper-title]. I retain only the right to distribute it for free for scholarly/scientific purposes, and in particular, the right to self-archive it publicly online on the Web.

Nevertheless, we would be happy to support departmental practices in this area provided that they achieve the same ends as we intend. The danger, however, is much more likely to be that departments are lax in this area, than that they are too stringent, even though the rights of the academic author to control both the unrefereed and refereed versions of their research articles are intuitively accepted. The fact is that many academics have not considered the implications of standing up to publishers in this matter. Academic colleagues testify to the fact that, where they have asserted their rights to secondary use of their own published material, publishers almost always back down. A university-wide policy, mediated by the library, will ensure that even timid researchers – such as those at the beginning of their careers who may be fearful of being rejected by publishers because of copyright assertions – will be able to find protection from a policy which applies across the university. In the final analysis, if a publisher is unwilling to acquiesce, the library can make a decision in consultation with academics in the relevant department about whether to accept unreasonable terms or not (and to archive a preprint plus corrigenda instead, as a legal substitute). But the brokering role then played by the library will help to expose recalcitrant publishers, and should enable more senior academics to come to the support of junior researchers in their departments in any war against bullying tactics employed by publishers – especially where these senior academics sit on editorial boards of the publisher in question. Indeed, the increased likelihood that someone will sit on such a board is one of the few benefits of the recent merging of publishing companies into a few hugely dominant companies. The library, in other words, as an independent administrator and standards-enforcer, can deliver a system in which best practice is consolidated and spread.

Lastly, as a neutral service within a university, the library can provide an institutional archive service which adheres to a common standard for bibliographic description and preservation. Leaving this work to individual departments to do would be to create an archive of very variable quality, particularly if description is left to academic authors, with the assistance of departmental secretaries. Since the Open Archive Metadata Harvesting protocol is a stringent one, the maintenance of a valid archive requires the rigorous consistency of approach which libraries have long shown to their own bibliographic databases (i.e. their catalogues).

The Conservative Tradition

Researchers do belong to a conservative tradition when it comes to publication. Winning them over to the adoption of an OAI-based self-archiving approach to scholarly publishing is far from easy. Among the other objections to change which we have encountered at Edinburgh are the following:

- There is considerable anxiety about the dilution of quality which will attend the ‘mixing’ of preprint with postprint submissions. The technical assurances on this are met with scepticism in some quarters.
- Disciplines which value postprint publication much more highly than preprint publication tend to favour the early release of publisher-copyrighted research articles onto public servers post-publication, seeing this as the ideal compromise. The economic argument for authors or institutions retaining copyright is not readily appreciated.
- There is some sympathy for learned society publishers, and a view of them as allies of researchers, which can dilute the force of the arguments against the practices of monopolistic, ‘profit-gouging’ publishers.
- Some academics fear an injunction from upon high within their institution which would force them to deposit all research into openly-accessible archives, thereby possibly jeopardising sponsorship conditions and providing too much visibility to

their own research. This can induce a very counter-productive sense of panic.

- Discipline communities within universities are generally not primarily concerned with the plight of other disciplines within the institution, or with the institution as a whole. Positive reasons to participate in the OAI also need to be advanced, and testbeds and demonstrators created so that academics can be reassured of the validity of the free access approach.

Because we know, from our experience to date, that convincing all our research-active academics in all faculties will be difficult and exhausting, we have adopted a two-pronged approach whereby we seek invitations to address groups of academic staff in departmental meetings, but at the same time meet with senior administrative staff. In general, the latter group have been very positive, and our pilot archive is making progress within the Faculty of Science & Engineering, with a number of articles identified for inclusion once the software is fully set up and hospitable to them. Our strategy will be to promote the archive widely across the faculty once it has around 50 or so articles in it, and we hope to encourage submissions from a wide constituency. Objectors will be referred to the authority we have established in securing the backing of the Director of Planning, Vice Principal for Academic Information Services, Chair of the Research Committee, University Librarian and other senior figures. We would hope that momentum will gather quickly once departments see their colleagues in other parts of the university engaged in the process, and some healthy competition between departments may yet prove of benefit. To this end, also, we are seeking funding from internal university sources to promote the archive across all faculties.

There is no prospect of coercion for departments which do not wish to join us in creating the archive. It has been suggested in some quarters that universities should require all researchers to post their CVs online, with links to their research outputs pointing to the institutional archive. This approach may be successful in the future, particularly as the submission date for the next RAE draws near, and especially if cautious departments face the prospect of having to locate all of their own outputs for the assessors, while those in the library scheme will have had the work done for them by the library in the course of their normal scholarly business.

Conclusions

Our early conclusions support the view that two conditions will need to be met before an institutional open research archive at Edinburgh gains widespread acceptance. Firstly, the rate of growth of compliant archives across the world will have to occur quickly. Academics will grow dubious of the utility of the initiative if they can find little of interest via an OAI service. At the present time, the beneficiary seems mainly to be the library, hoping to make savings in journal costs, though as the next RAE submission date approaches, benefits to departments should also become obvious. But hopefully by then many of our academics will be regular users of a wide free corpus of research publications via the OAI.

The other prerequisite is that the availability of research publication via the archives begins to permit savings to be realised in library budgets through cancellations of print and electronic journal subscriptions. This will lend impetus from the senior administration of the university to institutional and national efforts to increase the coverage of archives rapidly.

3 September 01.

© John MacColl. Non-exclusive right of publication granted to the European Conference on Digital Libraries, Darmstadt, 2001

Integrated Data Harvesting in the Perseus Digital Library System

David A. Smith
Anne Mahoney
Perseus Project, Tufts University
E-mail: {dasmith,amahoney}@perseus.tufts.edu

July 31, 2001

Abstract

The Perseus Project's document management system provides services for extracting information from diverse documents and reintegrating that information into document display. We demonstrate OAI services that integrate harvested metadata into the Perseus system and describe experiments with distributed information extraction on full text via the OAI protocol.

1 Introduction

Since its inception in 1987, the Perseus Project has reaped great benefits from a standards-based architecture. Despite the lack of tools for delivering structured texts to a wide audience, all Perseus' texts were encoded in SGML from the beginning, at first in an early version of the Text Encoding Initiative Guidelines [10]. This strategy paid off, most obviously, when in 1995 we were able to supplement a CD-ROM HyperCard delivery system with a World Wide Web HTML interface within a week.

In the past two years, we have been developing a document management system, known as the Hopper, for indexing and retrieving information in a variety of formats [9, 6]. Although the Hopper's native format is XML, other formats, such as SGML and PDF, are translated into well-formed XML for processing. By hiding the concrete markup elements (tags) from the application layer, the Hopper allows developers to build generic information extraction, visualization, and retrieval modules. Tight integration of digital library models supports not only top-down document retrieval but also lateral browsing strategies: the targets of one-way links point back to their sources, proper names and technical terms are automatically linked to reference works and other descriptions, and reading tools parse and define words in ancient languages. The Hopper is in production use on the Perseus website (www.perseus.tufts.edu), which receives

over a million page views, or about seventy thousand user sessions, every week. Perseus collaborators are currently testing installations of the Hopper with their own data, and we plan to have a public open source release of the system in late summer of 2001.¹

The Hopper takes the Resource Description Framework (RDF) as its metadata model and uses Dublin Core (DC) for most metadata semantics. We thus found it easy to become a registered Open Archives data provider, as well as a founding member of the Open Language Archives Community (OLAC); we expose metadata in OAI-standard unqualified Dublin Core and in the OLAC schema. We are now building on our integrated display environment to incorporate harvested metadata into document display and to experiment with harvesting mechanisms for sub-document information. In particular, we take the sites already running the Hopper as a prototype for fine-grained linking in a distributed digital library. We demonstrate the application of this library integration by linking names and terms in text, and different versions of the same text; linking words to dictionary definitions; reversing one-way links between documents; and gathering toponyms from disparate documents for geographic visualizations.

2 Integrating Harvested Metadata into the Hopper

When a reader ventures into an unfamiliar discipline, unknown names and terminology can hinder comprehension [4]. The Perseus digital library culls names and terms from the metadata of its documents and creates links in texts that use those words or phrases. Harvested metadata increases the set of terms to be linked: the titles, authors, and subject keywords of federated documents. When reading a speech by Demosthenes, for example, the user can click on the highlighted term “Arsopagus” and link to pictures of the site of this law court in Athens and also to an article in the Stoa Consortium about the history and procedures of the court (fig. 1). Note that although only the single term “Arsopagus” is linked, longer records whose titles only contain this term are also retrieved.

While linking authors, titles, and keywords is familiar from many genres of documents, from physics e-prints to Latin grammars, linking multiple versions of the same work is perhaps not such an obvious digital library function. In the terms of library science, this operation links different concrete *documents* to one abstract *work* [11, pp. 6–9] [5, pp. 26, 42]. For example, when viewing a translation of the *Iliad*, the reader can select other versions of the same document from a pop-up menu (fig. 2): another English translation, the original Greek text, and a facsimile edition at the Stoa site of a Greek manuscript edited by

¹Collaborators currently testing the hopper include the Max-Planck-Institut für the History of Science in Berlin, the Johns Hopkins Digital Knowledge Center, the Dillner Institute at MIT, and the Stoa Publishing Consortium (www.stoa.org); the last is already using the system on its production web server.

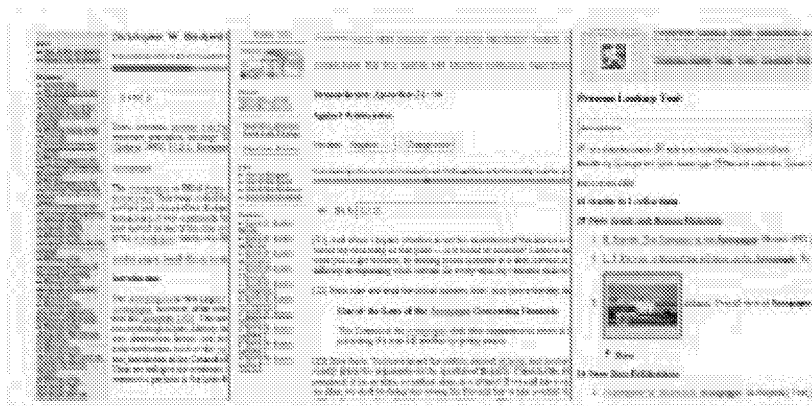


Figure 1: Linking from a highlighted term to an article in the Stoa Consortium. The Demosthenes in the center links from the term “Areopagus” to a list of resources, among them the article from the Stoa, on the left.

Domenico Comparetti. All of these documents are linked through the use of the Dublin Core element **Relation**, qualified by **IsVersionOf**. Although systems exist (e.g. [7]) for finding duplicate documents, and even parts of documents embedded inside larger ones, these systems of course require access to the full text, as well as being less effective when trying to link translations into different languages or text and image representations of a document. To overcome variants in authors and titles, catalogue systems often rely on some uniform title to provide the link. At present, the Perseus digital library is a *de facto* authority within the Hopper community for uniform abstract document titles; as the community evolves, however, we expect to formalize the maintenance of this and other authority lists.

As a member of the Open Language Archives Community, Persens is interested in linking together disparate resources, particularly for the study of historical and minority languages. Of particular use to the intermediate student trying to decipher a text is the ability in Persens to link to morphological and grammatical help and dictionary entries. Latin text in a Stoa publication, for example, can link to grammatical help and a dictionary in the Persens digital library (fig. 3). OLAC has defined a qualifier **Language** for the Dublin Core **Subject** to indicate that a particular resource describes the given language (as opposed to the DC **Language** element, which states the language in which a document is written). The Stoa site uses this metadata to make links to the dictionaries in Persens. Simply noting the presence of a Latin dictionary and linking all Latin words to it is not enough, however: the Renaissance Latin texts published by the Stoa contain many words not in the classically oriented Lewis and Short dictionary in Persens. Many links are thus made to non-existent dictionary entries. To give the user a better idea of what words can be successfully glossed, either the Persens data provider would have to expose all dictionary headwords as metadata (68,000 entries for the larger of Persens’

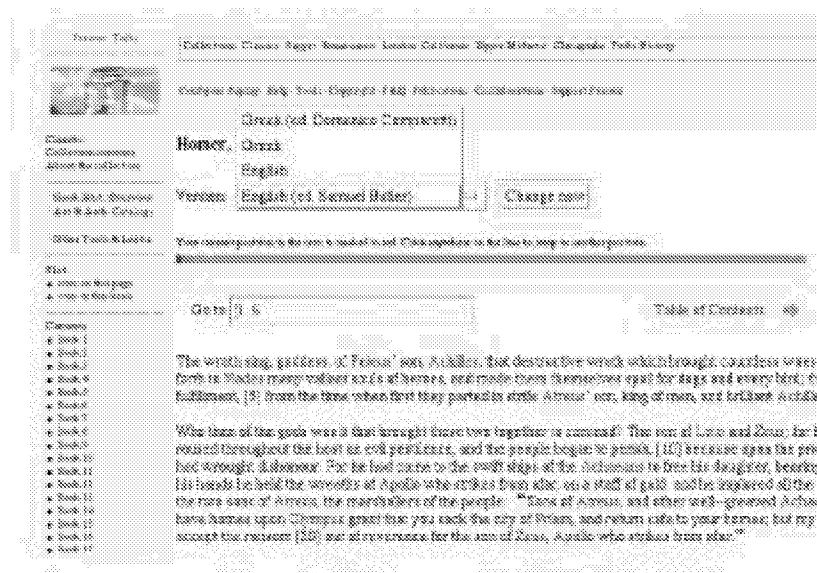


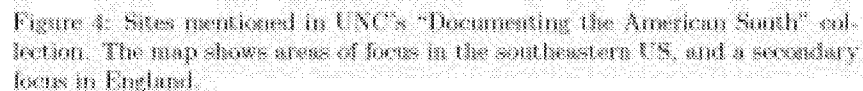
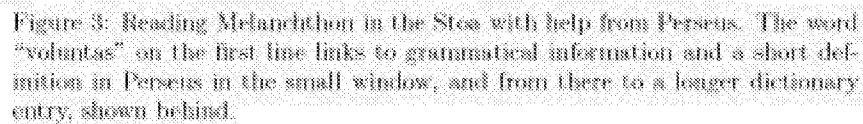
Figure 2: Linking several versions of the Iliad.

Latin dictionaries), or the Stoa would need some other mechanism to obtain this information.

Links on individual words are helpful for close language study, but higher-level visualizations can be useful for sensing the scope of a document or corpus. Since much data studied by both scientists and humanists has to do with locations on the earth, geographic visualizations are applicable across a wide variety of digital libraries. Whether assigned by cataloguers or directly extracted from a document, DC *Coverage* elements can be used to draw maps across distributed collections (fig. 4).

3 Distributed Information Extraction and Citability

Harvesting document metadata has proved to be a successful paradigm, especially where documents are about a single topic and relatively short. Although these conditions hold for most scholarly articles in e-print archives, the primary data, for example, on which such articles are based are not as tractable. First of all, primary sources are often quite long and amorphous and not about any single topic. (Is Stevenson's *Kidnapped* really *about* the 1745 uprising in Scotland?) Secondly, scholarly arguments often depend on citing specific passages, lines, or words in a source. Individual links from text span to text span — rather than just document to document — are thus desirable for following many chains of reasoning.



To address the need for sub-document metadata, we have taken advantage of the Hopper's ability to handle multiple DTDs and data formats to federate more detailed information across digital libraries. The harvesting Hopper pulls in not only Dublin Core and other metadata from the data provider but also the XML version of the document referenced by an `Identifier` element. Any number of the Hopper's information extraction and processing modules can then be run over the XML before it is discarded. Just as with other OAI service providers, sites can differentiate themselves by offering varying functionality. The Perseus digital library, for example, has developed named-entity recognition algorithms and a large gazetteer for finding and disambiguating place names in texts [8]. The Stoa or the Max-Planck-Institut need not acquire the gazetteer or run the algorithm, but its documents can still be harvested and mapped by the Perseus system, as in figure 4. The Johns Hopkins Digital Knowledge Center have developed their own full-text retrieval package; they can harvest the XML from Perseus and other libraries and index it for use with their searching package.

Citation extraction, or link analysis, is a service of interest to many digital libraries. The popularity of the ResearchIndex (formerly CiteSeer) citation linking service [2] and proposals for other citation and link databases described in [3] and [1] provide evidence for this need. As an example, see figure 5, where clicking on the italicized passage in the Latin poem of Propertius has brought up part of an article from the Stoa called "The Iconography of Amor in Propertius" that cites and quotes that passage. While the links a document makes could be exposed as metadata, this has some disadvantages. First of all, the context in which a link is made can carry important information about its usefulness. A citation made in the body of a text may be more relevant to the main argument than a link in a footnote; a citation that also quotes a passage from a text or occurs in the abstract may indicate an even closer connection between source and target. Also, as a practical matter, the citations in many heavily-used reference works take up more than half the space of the whole file.

It is also useful to be able to link to specific locations in documents, whether the user is following an explicit link or calling up the context of a search result. Search programs that merely link to the beginning of a two hundred page document whenever there is a hit in that document are not very useful. In order to produce consistent citations across multiple document formats, the Hopper metadata schema defines the repeatable element `Citation`, which contains a series of slots for generically citing a document. Most modern books and articles are cited by page alone, and thus have a single slot. For citing other kinds of materials, standardized chapters and sections, or acts, scenes, and lines are used. The Hopper's indices are able to convert these abstract citations into concrete locations in XML, for its own documents, or into page offsets in PDF or fragment identifiers in HTML (fig. 6).²

Practical issues will determine much of the future of distributed digital libraries. Many sites will not wish their full data to be harvested and indexed by external service providers, even if they are willing to provide large amounts of

²For the mechanism behind these mappings, see [9].

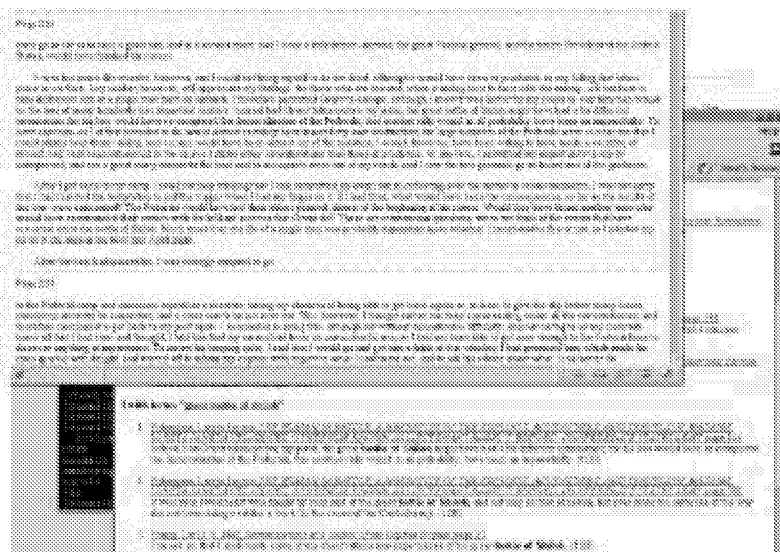


Figure 6: Search results in the lower window link to individual pages in the HTML document above, which is from a remote collection.

pre-extracted metadata. Many functions, such as title and keyword linking, can be performed across the entire Open Archives community. Others that involve locating and processing arbitrary passages in documents will require access to the documents themselves. With the public source release of the Hopper, and formalization of the Hopper community, we hope to build an even larger and more diverse testbed for research in distributed digital libraries and a community of service providers to make them useful.

References

- [1] Steven J. DeRose. XML linking. *ACM Computing Surveys*, 31(4), December 1999. http://www.cs.brown.edu/nemex/ACM_HypertextTestbed/papers/47.html.
- [2] C. L. Giles, K. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, pages 89–98, Pittsburgh, Pennsylvania, 1998.
- [3] Steve Hitchcock, Les Carr, Zhucan Jiao, Donna Bergmark, Wendy Hall, Carl Lagoze, and Stevan Harnad. Developing services for open eprint archives: Globalisation, integration and the impact of links. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pages 143–151, San Antonio, Texas, June 2000.

- [4] Steve Jones and Gordon Paynter. Topic-based browsing within a digital library using keyphrases. In *Proceedings of the 4th ACM Conference on Digital Libraries*, pages 114–121, Berkeley, CA, 11–14 August 1999.
- [5] S. R. Ranganathan. *Heading and Canons: Comparative Study of Five Catalogue Codes*. S. Viswanathan, Madras, 1955.
- [6] Jeffrey A. Rydberg-Cox, Robert F. Chavez, Anne Mahoney, David A. Smith, and Gregory R. Crane. Knowledge management in the Perseus digital library. *Ariadne*, 25, 2000. <http://www.ariadne.ac.uk/issue25/rydberg-cox/>.
- [7] Narayanan Shivakumar and Héctor García-Molina. Building a scalable and accurate copy detection mechanism. In *Proceedings of the 1st ACM International Conference on Digital Libraries*, pages 160–168, Bethesda, MD, 20–23 March 1996.
- [8] David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Proceedings of ECDL 2001*, page (forthcoming), Darmstadt, 4–9 September 2001.
- [9] David A. Smith, Anne Mahoney, and Jeffrey A. Rydberg-Cox. Management of XML documents in an integrated digital library. *Markup Languages: Theory and Practice*, 2(3), 2000.
- [10] C. M. Sperberg-McQueen and Lou Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative, May 1994.
- [11] Elaine Svenonius. *Intellectual Foundations of Information Organization*. The MIT Press, Cambridge, MA, 2000.

Heterogeneity in Open Archives Metadata

Gudrun Fischer, Norbert Fuhr
University of Dortmund, Computer Science 6
44221 Dortmund, Germany
{fischer,fuhr}@ls6.cs.uni-dortmund.de

July 10, 2001

1 Cyclades

The protocol specified by the Open Archives Initiative (OAI) [Van de Sompel & Lagoze 01] defines an easy-to-implement interface for harvesting metadata records. Any electronic archive or digital library that implements this interface is called an Open Archives data provider. The OAI interface by itself, however, does not provide any search functionality or user interface. To make the data from open archives available to end-users, further services have to be implemented on top of the Open Archives protocol. The Cyclades system is going to be one such service provider, implementing an environment for scholars and groups of scholars to search, browse, store, share and work on metadata records and other documents. In the Cyclades project, the University of Dortmund implements the metadata harvesting and indexing component, providing a uniform user interface for searching and browsing Open Archives metadata records.

In this paper, we discuss the current usage of the OAI specification regarding the metadata, and its consequences for a search and browse service. Based on empirical evidence from gathering 29 archives, we highlight different problems arising from different ways of using the OAI specification and discuss possible solutions.

The remainder of this paper is structured as follows: In the next section (2), we outline typical user requirements for sophisticated cross-archive searching and show in section 3 how they are partly met by the current OAI specification. In section 4, we outline several remaining problems which we believe can partly be solved by using Dublin Core qualifiers. In the last section (5), we summarize our findings and show further areas of heterogeneity in OAI metadata which cannot be dealt with by standardization alone and therefore should be treated by special OAI services.

2 Cross-archive searching and browsing

When searching and browsing across archives, a user will expect those search capacities that we also provided in a single archive environment. She will want to look for metadata records on documents that meet certain criteria, e.g. that belong to a certain author, or that date from a certain period of time. The

language of the document might be relevant, or the user might be interested in documents that contain certain keywords in the title or abstract. She might also want to browse a list of author names or languages in order to see which values are present at all etc.

In order to look for documents from, say, a certain period of time, the user should be able to formulate queries containing a comparison ("date before 2001-01-01 and date after 1999-12-31"). That implies that the dates contained in the metadata must be comparable, there must be a uniform date format and an ordering on that format.

Then, when the user is looking for an author named "Shakespeare", she is not interested in authors whose address is "Shakespeare St.", thus, if the author data can contain more than the name (address, affiliation), then the system has to distinguish between the author name and the rest of the author information. This distinction is also important for browsing: Metadata from the same single archive can be expected to have a uniform format for e.g. the author information, but it may vary greatly between archives. Thus, if the system cannot distinguish between author names and other data about the author, the lists it generates for browsing will probably not be of much use, entries beginning randomly with the author's affiliation, others with the author name, yet others maybe with the author's address.

Furthermore, when looking for documents written, for example, in English, the user will not want to bother with guessing the different keywords for "English" ("eng", "English", "en_us" etc.), she will just want to specify English as the document's language and leave the rest to the search service.

These examples show that it is not satisfactory to provide the user free text search only. Data typing (e.g. for dates and person names) is desirable, as well as uniform data formats in general, and controlled vocabularies when possible (e.g. for language).

3 Unqualified Dublin Core

One great advantage of the OAI protocol is the possibility to access any number of electronic archives in a uniform way and get records in at least one common metadata schema. All OAI compliant archives must provide Dublin Core metadata [Dublin Core Metadata Initiative 99]. This ensures that the user can explicitly search for different parts of document information, she can e.g. restrict her search for a specific author to those Dublin Core fields where an author can be expected (creator, contributor, publisher).

Dublin Core only defines a set of optional metadata fields, not restricting their content any further. Thus, the use of Dublin Core differs significantly among the individual archives.

Several archives list all the creators, or all the keywords, of one document in one single tag of the metadata record. This results in very heterogeneous browsing lists, since some list entries consist of e.g. one creator only (from archives that code each creator in a separate tag), while other list entries contain a whole list of creators. An additional problem arises when the individual values are not separated by delimiters, when they are listed in a single tag. As the definition of the Dublin Core elements states that e.g. a creator element should contain "An entity primarily responsible for making the content of the resource",

we suggest to follow the spirit of this specification and code each entity in a separate tag.

Although the date format is not specified in Dublin Core, the formats used in most of the archives we harvested comply with the ISO 8601 standard, however using different notations. The Dublin Core specification recommends to use the ISO 8106 [Wolf & Wicksteed 98] format YYYY-MM-DD for dates, respectively YYYY-MM-DDThh:mm:ssTZD for date and time (YYYY-MM-DD containing the date, the constant T separating the date and the time, hh:mm:ss specifying the time, and TZD being the time zone designator). This would ensure the comparability of date information, thus providing more search functionality than just free text, and still sparing the search service provider the need to write date conversion functions for the individual archives.

Unqualified Dublin Core is already a great improvement compared to having archives with arbitrary metadata schemes, as it allows to index and search different archives in a uniform way, and provides a minimum of semantics (given by the comments on the different fields in the Dublin Core specification). However, as it is, it still allows only for free text search, as the format and data types of the individual tags are not specified strictly. Since the Dublin Core specification recommends some formats as best use, however, we suggest to adhere to these recommendations when implementing an open archive data provider, as they already significantly improve the data quality with regard to searching and browsing.

4 Qualified Dublin Core

A means for specifying further details about the content of Dublin Core elements are *Dublin Core qualifiers* [Dublin Core Metadata Initiative 00].

Even if its format was standardized, the content of the date tag would still not always be comparable semantically. Does a date value specify the creation date of the corresponding document, the creation date of the metadata, the date when the metadata was entered into the archive, the date when the metadata was last changed? Without this additional information, a search for documents e.g. "from the year 2000" will probably result in numerous false hits. This type of problem can be solved by using Dublin Core *element refinement qualifiers*. These qualifiers help to define the semantics of a tag more closely, marking a date value as creation, validity, availability, or modification date, respectively.

Without further restrictions on the tag contents, as is now the situation in the OAI specification, e.g. the content of the language tag for English documents may range from "en,gb" over "English" to "19th century English with passages in French". To make information about languages or the subjects of documents more comparable, Dublin Core recommends to use controlled vocabularies. These can be explicitly referred to by *encoding scheme qualifiers*.

This kind of qualifiers can also be used to define a data format or parsing rules for the content of a tag, thus allowing a search and browse service to recognize significant subcomponents of an element value.

However, there are yet some Dublin Core elements, where there is no recommended standard format, e.g. the creator, contributor and publisher tags. Some archives, for example, obviously list authors (in the tags *creator*, *contributor*, and sometimes *publisher*) with their affiliation, but without any separator

indicating where the author name ends and where the affiliation information starts. Without further standardization, these elements remain a source of false hits and confusing heterogeneity while browsing.

5 Conclusion and Outlook

In this paper, we have mentioned mostly syntactic differences between different open archives. In order to provide meaningful search and browse services, this heterogeneity needs to be reduced. We therefore strongly recommend using the Dublin Core elements according to the respective recommendations in the Dublin Core specification. Additionally, Dublin Core qualifiers would significantly improve the comparability, parsing, and interpretation of the data. In order to cover all Dublin Core fields and treat substructures, additional Dublin Core qualifiers need yet to be developed.

Still, there remain differences between archives that cannot or should not be resolved by standardization, e. g. the use of different languages in the metadata, or the use of different classification schemes and controlled vocabularies. It is surely not possible to define one vocabulary that is valid and usable for all metadata records that could ever exist. Neither is it desirable to require that metadata be coded in one single language only. Thus, this kind of heterogeneity cannot be avoided and should be dealt with by special mapping and interoperability services. These are currently a subject of investigation at the University of Dortmund.

References

- Dublin Core Metadata Initiative. (1999). *Dublin Core Metadata Element Set, Version 1.1*. <http://dublincore.org/documents/1999/07/02/dces/>.
- Dublin Core Metadata Initiative. (2000). *Dublin Core Qualifiers*. <http://dublincore.org/documents/2000/07/11/dces-qualifiers/>.
- Van de Sompel, H.; Lagoze, C. (2001). *The Open Archives Initiative Protocol for Metadata Harvesting, Version 1.1*. <http://www.openarchives.org/OAI/openarchivesprotocol.htm>.
- Woff, M.; Wicksteed, C. (1998). *Date and Time Formats*. <http://www.w3.org/TR/NOTE-datetime>.

Enhancing OAI Metadata for Eprint Services: two proposals

Tim Brody, Zhunan Jiao, Steve Hitchcock, Les Carr and Stevan Harnad
*Open Citation Project, IAM Research Group, Department of Electronics and Computer
Science, University of Southampton SO17 1BJ, UK*
Contact email: sh94r@ecs.soton.ac.uk

The Open Archives Initiative has always maintained a distinction between data providers and service providers. This works at a functional level; some current projects are showing that it is less significant at an operational level. The Open Citation project harvests reference data from full-text eprint archives for reference linking, citation analysis and citation-ranked search. These are regarded as service provider functions where the end-user interacts directly with the services. The aim is to make these data available for export back to the full-text archives and to other service providers. Thus OpCit becomes a data provider too. The current Open Archives Protocol for Metadata Harvesting says nothing about full-text data harvesting for services such as these, nor about the export of processed data. This short paper outlines two proposals for progress on these issues.

1 Introduction

Emerging Open Archives services are blurring the distinction between data and service providers introduced in the original Santa Fe Convention framework documents describing the Open Archives Initiative (OAI). (Van de Sompel and Lagoze 2000) This is highlighted by the recently announced Kepler framework, a broker-based peer-to-peer network architecture, such as that used by Napster, which provides individual authors with software to set up personal archives, hosts a registration server and harvests data for subsequent dissemination. (Maly *et al.* 2001) This blurring between provider functions has significant practical implications for OAI.

The OAI Protocol for Metadata Harvesting (Van de Sompel and Lagoze 2001) mandates simple, Dublin Core based metadata to achieve interoperability between archives with low overhead for archive maintainers. This focus on simplicity appears to be vindicated by the apparent demise of NCSTRL, a forerunner of OAI as a collection of distributed archives. (Krichel and Warner 2001b) It suggests, however, the expectation of data transfer between services at a fairly low level of functionality unless OAI data and service providers can supplement the basic metadata for particular application areas, as allowed in the OAI protocol.

An example of enhanced metadata is the proposed Academic Metadata Format (AMF), a parallel metadata set that can be deployed with basic OAI metadata and which is designed to be used by the eprint archiving community, or more generally 'to advance scholarly communication over the Internet'. (Krichel and Warner 2001a)

This is a welcome development. Once an OAI service provider itself becomes a data provider, it is inevitable that data output in this case will not be simple document metadata as offered by the original document archives (which might be a better description than 'data provider'), otherwise, what service would the service provider be offering?

This paper considers the enhancement of OAI metadata from the perspective of the Open Citation (OpCit) project, which is demonstrating reference linking and citation analysis services by working with the largest OAI-compliant archive, the Los Alamos physics archives (arXiv). The approach described is intended to be generalised for other Open Archives, and is presented here to promote discussion and participation in the process.

2 Reference linking and citation analysis

To provide reference linking and citation analysis requires more data than is included in the OAI metadata. In principle, the full reference list is required (the full text is required for in-text, context reference linking, i.e. linking the occurrence of the citation within the text to the reference at the end of the text). OAI metadata does not provide mechanisms to expose and harvest full content. (Warner 2001)

An early OpCit reference linking demonstrator was described by Hinchcock *et al.* (2000). The reference database continues to be kept-up-to-date and has since been restructured to store richer data, improving accuracy and robustness. This has enabled the service to be extended, providing forward (in time) citation links as well as a Google-like search service that ranks results according to citations or hits.

The project has achieved this by working in partnership with the arXiv maintainers. To extend this approach to other archives, the project confronts two questions:

1. How can reference lists be extracted from Open Archives within the framework of the OAI?
2. How can processed data be exported back to the original archives so that it can be visible to the users of the archives (rather than as a standalone demonstrator)?

Since OAI data is intended to be interoperable, and OAI service providers are envisaged as cooperative, it is reasonable to assume that other services, search engines or Web portals for example, not simply the original archives, may also want to import the processed data. Thus a generalised interface to expose the data is desirable.

In the OpCit application the schematic in Figure 1 shows data input and output in the context of the OAI. The only parts of this schematic mandated by the OAI are the target paper archives and OAI metadata output. OpCit is responsible for the citation database and specified user services. The subjects of the questions above are the production of the reference list, and data export (stages A and D, respectively, in the schematic).

2.1 Extracting and parsing reference lists

Experience of large archives shows that document maintainers are unwilling to permit automated downloading of full texts. Alternatively, OpCit has developed software to identify and extract reference lists from papers, and this is available for free download and use by archive maintainers to create a separate collection of reference list documents, saving the main document server from possible overload.

This series of Perl modules is available from <http://ambica.ees.soton.ac.uk/code/doc/ReadMe.html>, and includes:

- Markup_TeX.pm: inserts 'xxxOpCit' at the beginning of each reference in the TeX source file. This mark-up is used by 'Parser_DVI.pm' to identify each reference;
- TeX2DVI.pm: converts TeX/LaTeX to a DVI file, then DVI to text by 'dvitype' (Unix command);
- Parse_DVI.pm: parses the text file created by 'dvitype' to produce a list of references;
- Citation.pm: parses each reference (citation) string to discover its metadata (authors, journal, volume, issue, etc.)

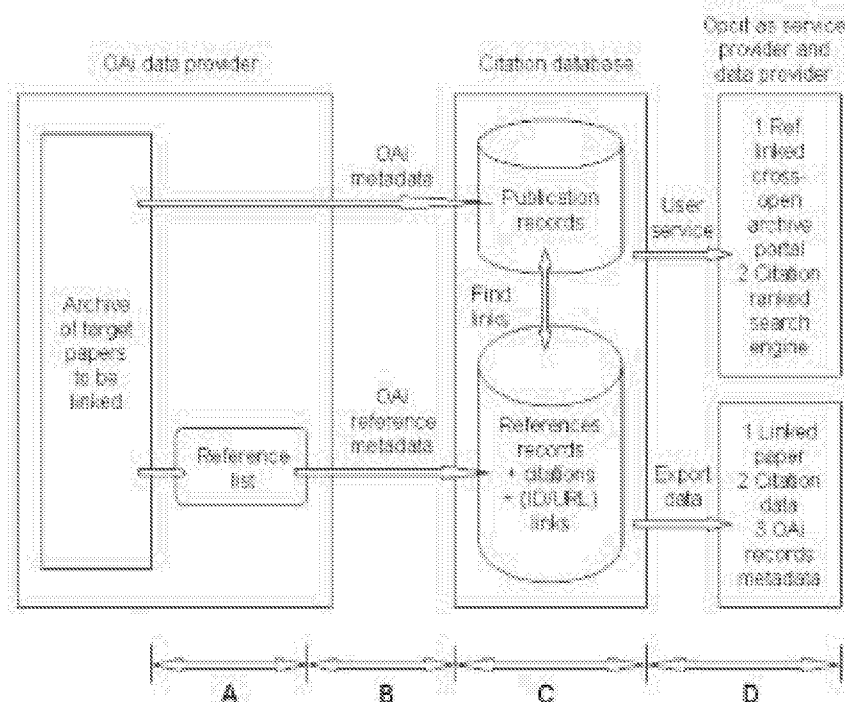


Figure 1. Proposed schematic of data input and output from the OpCit citation database

This software adds little manual overhead to maintenance beyond initial set-up, and imposes no requirements on authors. Initially these modules are optimised for arXiv because of the common TeX format found there, although other versions have been used with pdf and html. Perl scripts that make calls to the above modules can be developed locally.

2.2 Citation database

The parsed reference data extracted using these scripts are stored in the classic citation database structure shown in Figure 1 along with the conventionally harvested OAI metadata. Comparing the reference records with the publication records for an archive enables links to point at those referenced documents held in the archive. Also, for each publication record it is possible to use the reference records to determine if it has been cited. This can be displayed in a number of ways, as shown in Figure 2.

Within the project this database, called cite-base, is also used to serve the cite-baseSearch engine (<http://cite-base.ecs.soton.ac.uk/cgi-bin/search>) that ranks results according to citations or hits, as selected by the user.

2.3 Data export

While the project is capable of managing data capture, database maintenance and the user interface within the current framework, a more intriguing prospect is exporting data, principally back to the archive maintainers so that the services demonstrated above can be made available to users of the archive, but also to other OAI service providers.

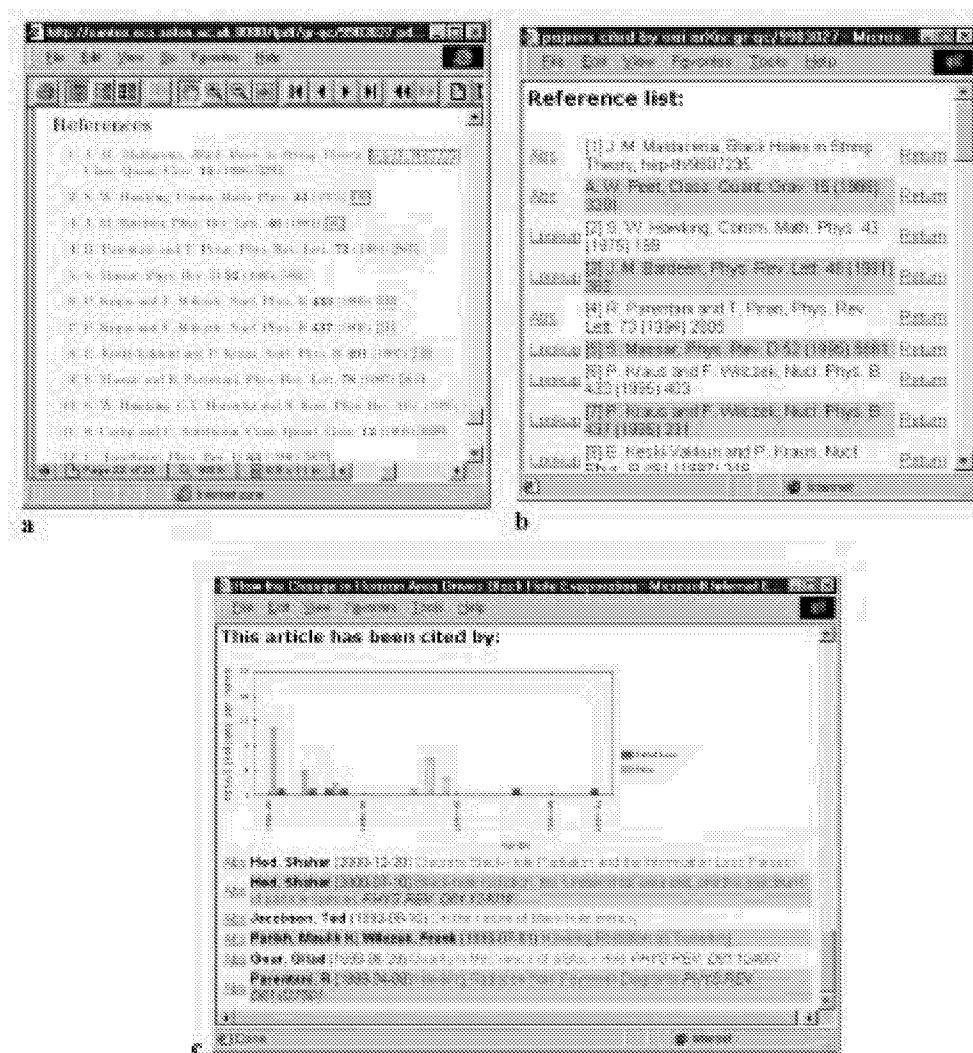


Figure 2. Reference and citation lists for a paper in the physics archives: **a**, reference links added to the original pdf; **b**, the same reference list extracted from cite-base with direct links (Abs) to texts and Lookup links to search for e.g. other works by the cited authors where the cited document is not available; **c**, citations to the paper from which the reference list was taken.

OpCir's immediate requirement is for a metadata format to express the reference lists (and hence the referenced articles), along with fields such as title, abstract, etc. This can be expressed using Dublin Core metadata, specifically the *relation* attribute. An in-house format, "opcit_dc" (Figure 3), augmenting Dublin Core with a *citation* attribute, was developed to store structured data for references (title, author, year of publication). *Citation* allows the reference and linked identifier to be expressed (with hindsight this is probably possible using Dublin Core qualifiers, without having to modify the schema).

Two proposals for the OAI community to consider emerge from this brief description of the reference linking and citation analysis work of the OpCit project:

- Richer formats are required to supplement the basic OAI metadata and to expose data for transfer between service providers and archive maintainers. The Academic Metadata Format appears to be a good foundation for this, and others are encouraged to participate in the development and review of AMF so that it might be adopted with some consensus.
- Reference lists need to be available for automated download from archives. Modular software aimed at archive maintainers, which automates the extraction and mark-up of references from submitted papers, is available for free download from the OpCit project.

References

Hitchcock, Steve, *et al.* (2000) Developing Services for Open Eprint Archives: Globalisation, Integration and the Impact of Links. *Proceedings of the Fifth ACM Conference on Digital Libraries* (ACM: New York), 143-151, June. Version available at <http://opcit.eprints.org/dl00/dl00.html>

Krichel, Thomas and Warner, Simeon M. (2001a) A Metadata Framework to Support Scholarly Communication. Submitted to the *Dublin Core Conference*, Japan, October. Draft available at <http://openlib.org/home/krichel/kanda.html>

Krichel, Thomas and Warner, Simeon M. (2001b) Disintermediation of Academic Publishing through the Internet: an Intermediate Report from the Front Line. *ICCC/IFIP Conference on Electronic Publishing*, Canterbury, UK, July. <http://openlib.org/home/krichel/sants.html>

Liu, Xiaoming, *et al.* (2001) Arc - An OAI Service Provider for Digital Library Federation. *D-Lib Magazine*, Vol. 7, No. 4, April. <http://www.dlib.org/dlib/april01/liu/04liu.html>

Maly, Kurt, Zubair, Mohammad and Liu, Xiaoming (2001) Kepler - An OAI Data/Service Provider for the Individual Access. *D-Lib Magazine*, Vol. 7, No. 4, April. <http://www.dlib.org/dlib/april01/maly/04maly.html>

Van de Sompel, Herbert and Lagoze, Carl (eds) (2001) The Open Archives Initiative Protocol for Metadata Harvesting. Protocol Version 1.1, 2 July. <http://www.openarchives.org/OAI/openarchivesprotocol.htm>

Van de Sompel, Herbert and Lagoze, Carl (2000) The Santa Fe Convention of the Open Archives Initiative. *D-Lib Magazine*, Vol. 6, No. 2, February. <http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>

Warner, Simeon (2001) Exposing and Harvesting Metadata Using the OAI Metadata Harvesting Protocol: A Tutorial. *HEP Libraries Webzine*, No. 4, June. <http://library.corn.ch/HEPLW/4/papers/3/>

Integration of grey literature with electronic journals in the CERN Library

Workshop
Experimental OAI-based
Digital Library Systems

8 September 2001
ECOL, Darmstadt, Germany

Corrado Pettenati
CERN (ETT-SI)

Presentation plan

- The CERN Library
- Definitions
- CERN grey literature management
- OAI protocol at CERN Library
- A proposal for a new service in OAI protocol
- Conclusion

5/4/02

2



The CERN Library

- A central unit and four satellites
- Few monographs and conference proceedings, fewer than 40,000
- Subscriptions to scientific journals
 - 1200 titles available electronically in full text
 - 450 titles of those also available in paper edition
- A very large collection of grey literature, more than 430,000 documents
 - Half of them with full text electronically available from February 1994 onwards

5/4/02

3



Definitions

The CERN grey literature collection is mainly composed of

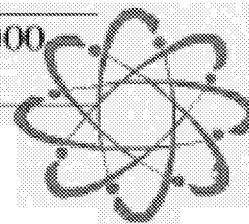
- Documents prepared to be submitted to scientific journals (Preprints and Scientific Notes)
- Documents submitted to conferences
- Theses

5/4/02

4

Number of accesses to the CERN Library catalogue

| | | |
|---------------------------------|----------------------------------|------------------------------|
| Access per day to the catalogue | Access per day to the f-t server | Documents downloaded per day |
| 25,000 | 16,000 | 2,000 |
| Pictures downloaded per day | Ratio internal vs. external use | Distinct hosts per month |
| 450 | 30:70 | 20,000 |



5/4/02

Community

Users are:

- Physicists at CERN and all over the world
- Distinct hosts counted in 2000:
 - Total of 127 000 distinct hosts
 - 8 000 at CERN
 - 93 000 outside CERN
 - (26 000 unresolved IPs)
- On average, 20,000 distinct hosts per month

5/4/02

6

CERN grey literature

procedures and management

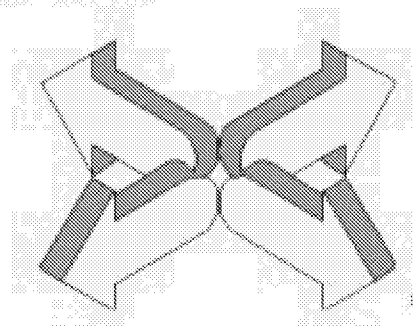
- Direct electronic submissions
 - Official series
 - Open series
 - Theses
- Downloading from other grey literature servers
 - Los Alamos, DESY, SLAC, FNAL, Dubna, INFN, etc.
- Digitization of paper documents
- Exchange with other labs (Annual reports)
- Harmonization of the record description

5/4/02

7

Provenance

- More than 50,000 documents processed per year
 - Internal to CERN 10%
 - External 90%



5/4/02

8

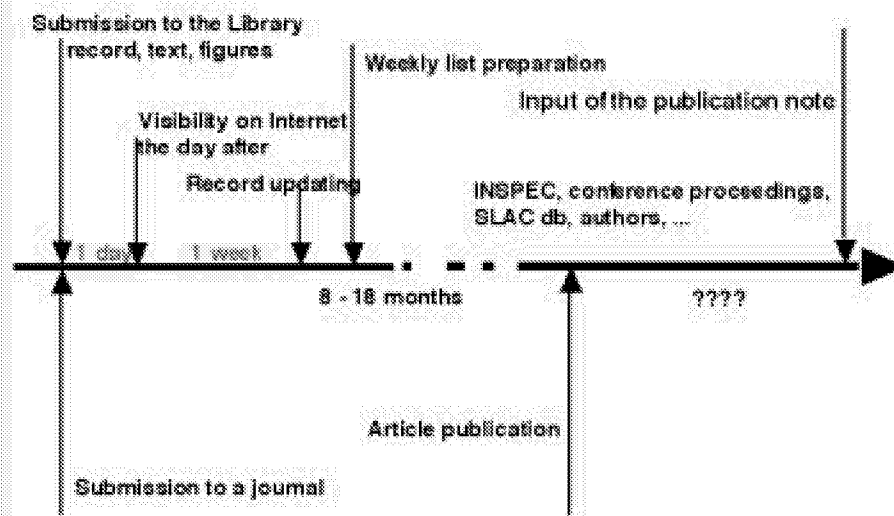
Documents prepared for publication: Preprints

- They are sent to the CERN Library and at the same time submitted to the publisher of a scientific journal
- They are distributed via the Library Web server the day after submission
- In general they will be published much later, *after 8-24 months*

5/4/02

9

Preprints processing procedure



5/4/02

10

SYN0-0251285

BA 11

LM eng

YR 1997

SW dds n 88w 9725

ER hep-th/9706105

EM Thomas.Schucker@cpt.univ-mrs.fr

TI Noncommutative Yang-Mills and Noncommutative Relativity : A Bridge Over Trouble Water

AU2 Carminati, L.

AU Iochum, B.

AU Schucker, T.

AF Marseille

IM dds 16 Jun 1997 dds 29 p

EKT dds http://preprints.cern.ch/cgi-bin/setlink?base=preprint&categ=hep-th&id=9706105 dds

Access to full text document

IM CPT-96-P-3477

SU AD PARTICLE PHYSICS:THEORY

SUT AD

AB Connes' view at Yang-Mills theories is reviewed with special emphasis on the gauge invariant scalar product. This landscape is shown to contain Chamseddine and Connes' noncommutative extension of

AB dds general relativity restricted to flat space-time if the top mass is between 172 and 204 GeV. Then the Higgs mass is between 188 and 201 GeV.

YAU Carminati, Lionel

YAU Iochum, Bruno

YAU Schucker, Thomas

5/4/02

11

Title : Noncommutative Yang-Mills and Noncommutative Relativity: A Bridge Over Trouble Water

Author : Carminati, L.

Imprint : 16 Jun 1997 - 29 p

Server : hep-th/9706105

Report no : CPT-96-P-3477

Abstract : Connes' view at Yang-Mills theories is reviewed with special emphasis on the gauge invariant scalar product. This landscape is shown to contain Chamseddine and Connes' noncommutative extension of

URL address : http://preprints.cern.ch/cgi-bin/setlink?base=preprint&categ=hep-th&id=9706105

Add.author : Iochum, B., Schucker, T.

Lib Coll Shelf number

C SVR hep-th/9706105

Vol. text

000

Category

08

5/4/02

12

Comes' were being sold. The scene is captured with special emphasis on the young woman's outfit. This landscape is shown to contain Chamek's and Comes' noncommittal extension of general relativity restricted to flat space-time, if the top mass is between 177 and 204 GeV and the Higgs mass is between 130 and 201 GeV.



Preprint (published preprint) (cond-mat/930805)

Anomalous Behavior at a Superconducting Quantum Critical Point

by Ramazashvili, R ;

Email: Ramazashvili@physics.utdallas.edu

Publ. Ref.: Phys. Rev. B : 58 (1997) 5518 - Published version :

Imprint: 4 Nov 1997 4 p

Subjects: CONDENSED MATTER

Abstract:

Motivated by pressure experiments on $\text{UBa}_2(12)$ and $\text{N}_2(2)\text{Ba}_2\text{CuO}_6$, we discuss low-temperature effects of the pairing interaction at a superconducting quantum critical point in a clean system. We find that measurements at this quantum critical point can provide a diagnostic tool to mark out non-BCS mechanisms of superconductivity.

[Access to fulltext document](#) : [Show references](#) : [Cited by \(experimental\)](#) : [Mark document](#)

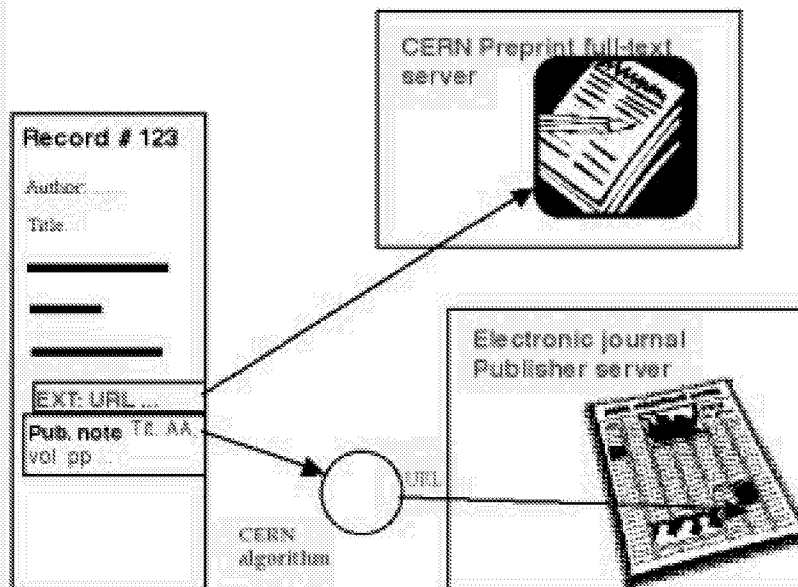
[Home](#) : [Comments](#) : [Help](#) : [Show Shelf](#) : [Format Output](#) : [Check Loans](#) : [New Searches](#)

Access to the
preprint full text

Access to the published
text

5/4/02

15



5/4/02

16



The triplet (title, vol./year, pp)

- Very precious
 - Because it works also for papers precedent to the "electronic age"
 - *Publishers are making their retrospective collections electronically available*
- Simple and intuitive to use
 - The data triplet is more intuitive than a DOI
- ... and it works with APS, AIP, IOP, Baltzer, Springer, SIF, EDPSciences and (partially, only NPE service) Elsevier

5/4/02

17



Document formats

Accepted

- Tex/Latex
- Word
- TIFF
- HTML
- ...

Distributed

- PDF
- PS
- HTML
- TIFF
- GIFF
- ...

5/4/02

18

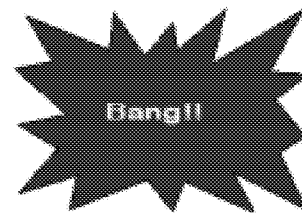
Electronic submission processing

- Conversion from Tex/Latex to PS
- Conversion from Word to PS
- Conversion from PS to PDF
- Preparation for the full-text searching and citation processing

5/4/02

19

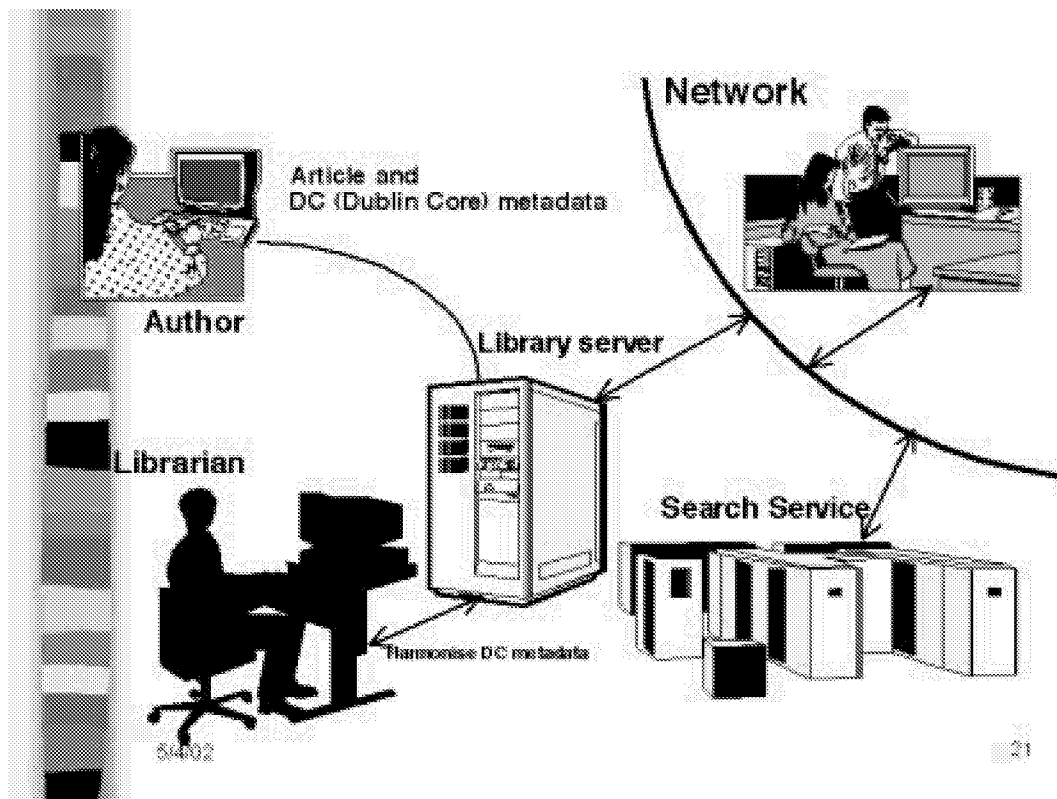
Citations management



- The document in PDF is analysed and citations are automatically extracted
- If the cited document is also in the CERN database a *hard-link* is inserted next to the citation
- If a "triplet" can be defined, data are set up for calculation "on-the-fly" at the display moment, of the link to the cited article
- The citations can not always be safely processed automatically; 85-90% correctly automatically processed

5/4/02

20



21

OAI protocol

- Implementation at CERN Library
- Current status and planning
 - Information from J.-Y. Le Meur (ETT-DH)

5402

22

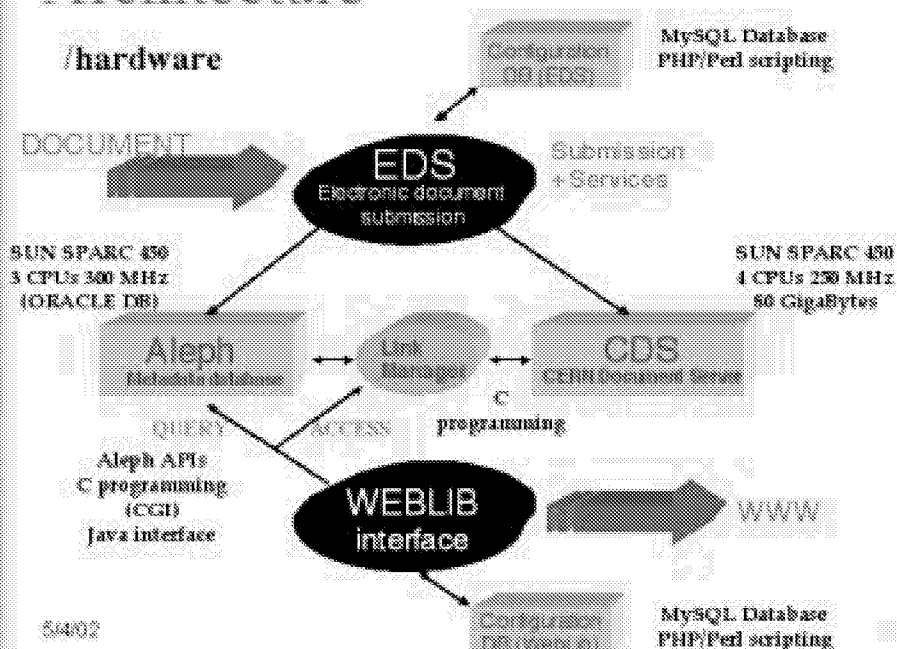
Background: CERN Library database

- Aleph 330 library system (Ex-Libris)
- Customized Web interface: WebLib
- Software built on top of Aleph APIs (RPC)
- Two main servers: *ALEPH* and *CDS*
- A separate MySQL database for 'non library' documents

5/4/02

23

Architecture



5/4/02

24

OAI @ CERN: history

Metadata acquisition (since 1994)

- Manual: collection of *scanned* documents
- Electronic:
 - Web & email *submission mechanism*
 - *Uploader* application for metadata transformation
 - Checked by human
- Long term storage system with an open interface for collecting the metadata

Involvement in OAI (1999)

- Close follow-up since Santa Fe meeting
- Straightforward objectives for CERN:
 - Metadata exchange simplification
 - Metadata proofreading savings

5/4/02

25

OAI 1.0 @ CERN: status

A test collection:

- composed of books and eprints
- 300,000 records extracted from our Library system
- Stored in a MySQL database (based on MARC 21)

OAI 1.0 compliant with:

- Three formats supported: *oai_dc*, *oai_marc* and *oai_rfc1807*
- All functions implemented: *Identify*, *ListSets*, *ListMetadataFormats*, *GetRecord*, *ListIdentifiers*, *ListRecords*
- *oai:cerncds:xxxx* ready but not in production yet

5/4/02

26

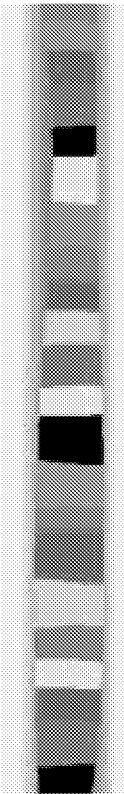


Waiting for ALEPH500

- ALEPH500 is OAi-compliant
- To avoid reorganizing our library automation services twice in a short period of time
 - We will first implement ALEPH500 (*before end of the year*)
 - Then we will convert our database from ALEPH330 to ALEPH500
 - This implies also a metadata format conversion
 - Then we will open the CERN OAi data and service providers

5/4/02

27



Proposal for a new service in OAi

- Suggestions presented by J.-Y. Le Meur in Berlin at the European OAi Workshop at the end of February 2001
- Suggestions presented by T. Baron in Geneva at the OAi Workshop at the end of March 2001

5/4/02

28

General Issues

Harvester distinction?

- Kind of "OAI Intranet" would be useful
- Different sets for different partners?

OpenURL in OAI?

- OAI format already as a Web output format in our test collection (e.g. search by author and give OAI output)
- Agreed protocol necessary for searching many OAI compliant sites in parallel

Full-text data provider within OAI?

- Full text exchange with agreed protocol

Increase metadata quality?

- Too few mandatory tags in DC
- Specific tags agreed for specific communities

5/4/02

29



Validation and OAI

- CDS is ready for OAI compliancy as data provider
- In OAI philosophy: document quality is not recorded
- How to keep the value added by the validation?
- Simple solution: adding a quality label
 - Set-wide
 - Record-specific

OAI and peer review Workshop
(CERN 22.03.2001)

Dimitris Boneas / Peter Strobel

5/4/02

30



Can we afford to lose the validation information?

<http://cds.cern.ch>

OAI and post review Workshop
JCERN 22/03/2003

Thomas Hain - CERN

31

Another OAi service

As long as the current publishing framework is in use

- We should be interested in inputting, maintaining and spreading the publication note in a metadata field with three subfields
 - Title of the journal
 - Volume or year
 - Pagination
- The publication note should be able to generate automatically the access URL to the published version of the document
- The pointer to the preprint full text should remain for the benefit of those readers without a licence to the journal

5/4/02

32

The triplet (title, vol./year, pp)

- Very precious

- Because it works also for papers precedent to the "electronic age"

- *Publishers are making their retrospective collections electronically available*

- Simple and intuitive to use

- The data triplet is more intuitive than a DOI

- ... and it works with APS, AIP, IOP, Baltzer, Springer, SIF, EDPSciences and (partially, only NPE service) Elsevier

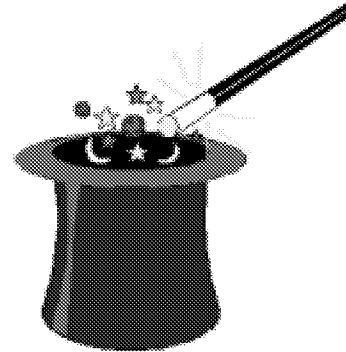
5/4/02

33

Future

Short term

- CERN as data provider
 - ... for CERN-specific collections
- CERN as data harvester (and service provider)
 - setting up a data harvester
 - enabling extended service (e.g. cross-archive searches, personal baskets and alerts)
 - ... for High Energy Physics readership



5/4/02

34



Far future

Long-term hopes

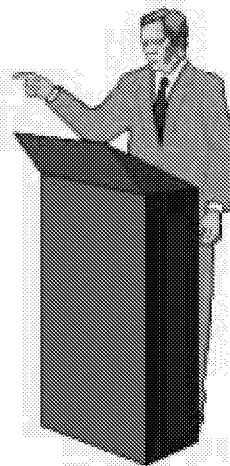
- All HEP institutes OAi compliant
... for metadata AND data
- Parallel searching possible (with OpenURL protocol)
- OAi also used inside CERN between various applications (Engineering Database, Administrative Documents, HEP Preprints, ...) to build the CERN long-term electronic archive

5/4/02

35

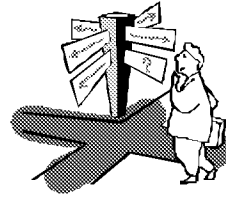


QUESTIONS?



5/4/02

36



The Use of Open Archives

Who, How Often and Why

Les Carr

University of Southampton

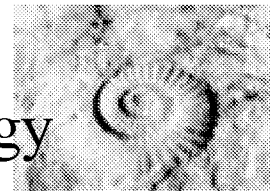
Overview

- Reasons for Introspection
- Usage analysis
 - depositors
 - readers
- Attitude Analysis
- Reflection

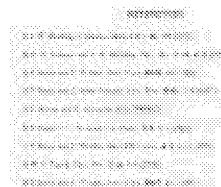
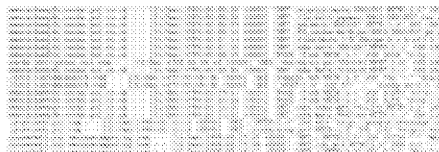
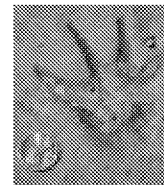
Salutary Warning

- A scholar is just a library's way of making another library
 - *Daniel Dennet, Consciousness Explained*

Archive Palaeontology



- Palaeontology tells us about what really happened in the past by examining its preserved remains
 - articles
 - metadata
 - weblogs



Reasons for Introspection

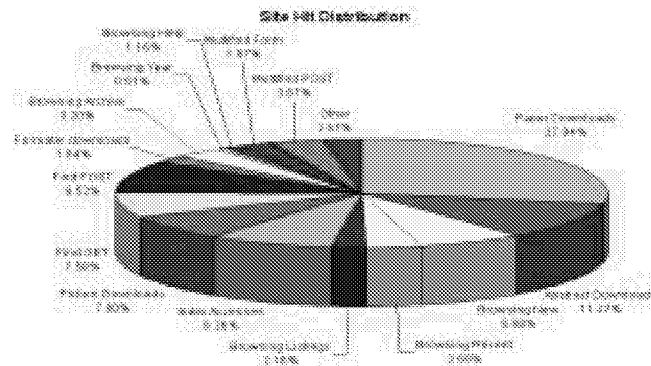
- “...metadata are expensive to create - it is estimated that tagging papers with even minimal metadata can add as much as 40% to costs...”
- “...not all papers will warrant the costs of marking up with metadata, nor will much of the grey literature, such as conference proceedings ...”

– *Declan Butler, Nature, 06 September 2001*

Reasonable Questions

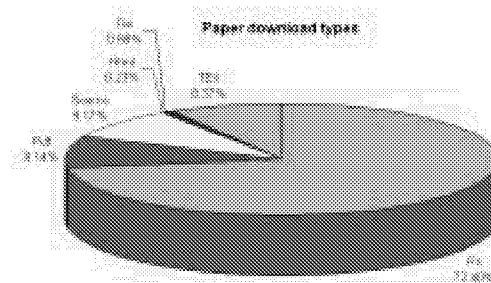
- Is the grey literature worthless?
- Or is it identifiably embryonic stages of valuable scientific communications?
- If so, what are e-scientists looking for?
- What metadata can we add to help them?

What Do Users Look At?



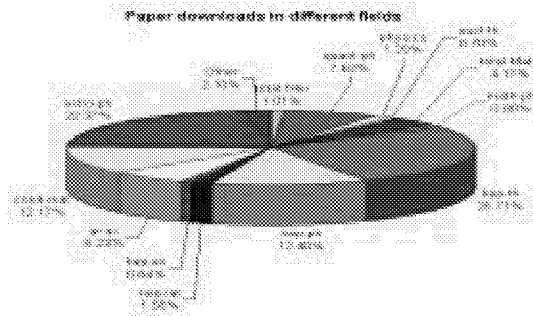
- articles 28%, abstracts 11%
- search 23%, content listings 13%

What Article Formats?



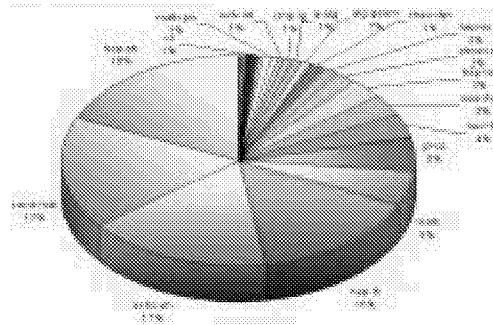
- Mainly PostScript (not necessarily for printing)

Archive Areas: Reading



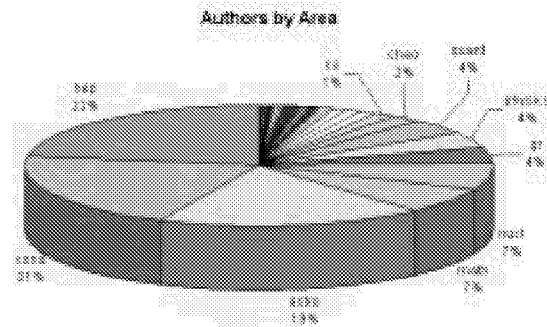
- Archive is many different sub-areas
 - Often with different behaviours
 - No such thing as “a Physicist”

Archive Areas: Depositing



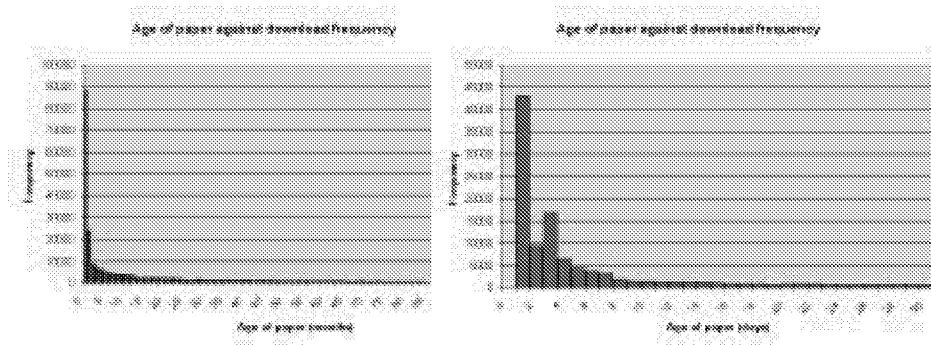
- Archive is many different sub-areas
 - Often with different behaviours
 - No such thing as “a Physicist”

Archive Areas: Authors



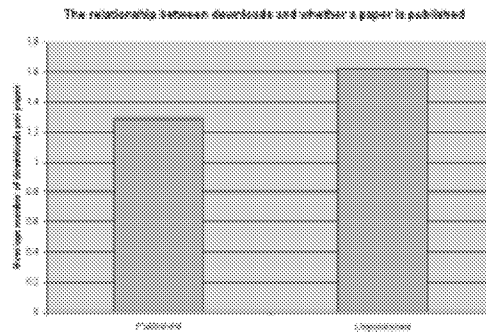
- Archive is many different sub-areas
 - Often with different behaviours
 - No such thing as “a Physicist”

Article Preferences: Age



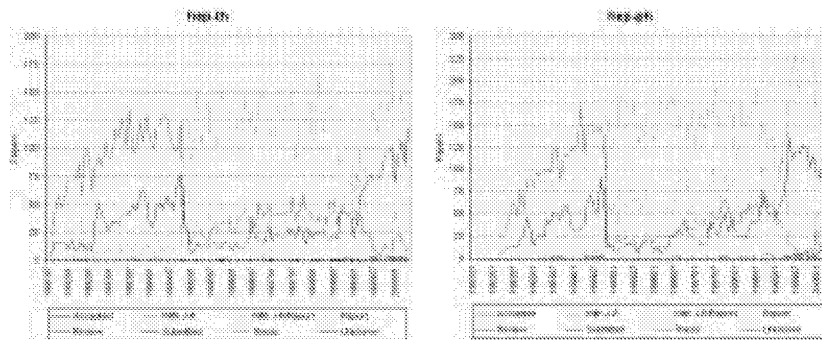
- Is this contrary to expectation?

Article Preference: Status



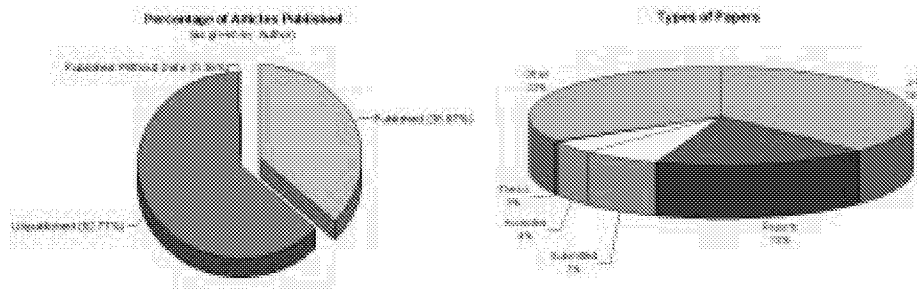
- Do we approve of this?
 - See Harnad on Sky-Writing.

Publication Lag



- Unpublished status crosses over with published status about a year before 'now'

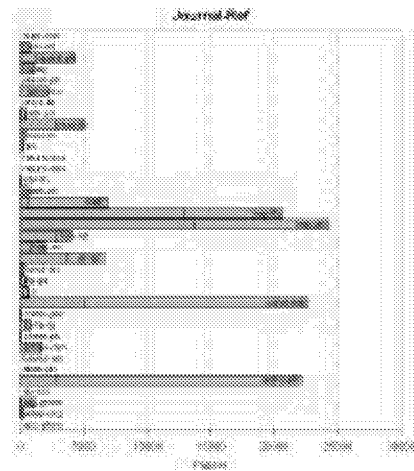
Do EPrints Get Published?



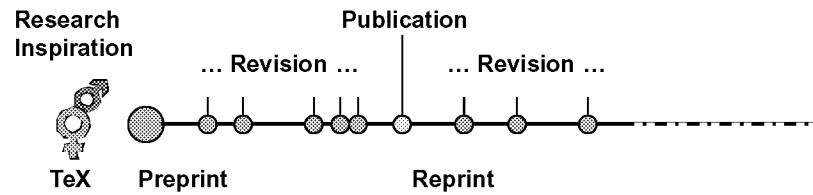
- At any one time, less than 50% seem to be 'published'

Do EPrints Get Published Equally?

- Proportion of publication varies across archive

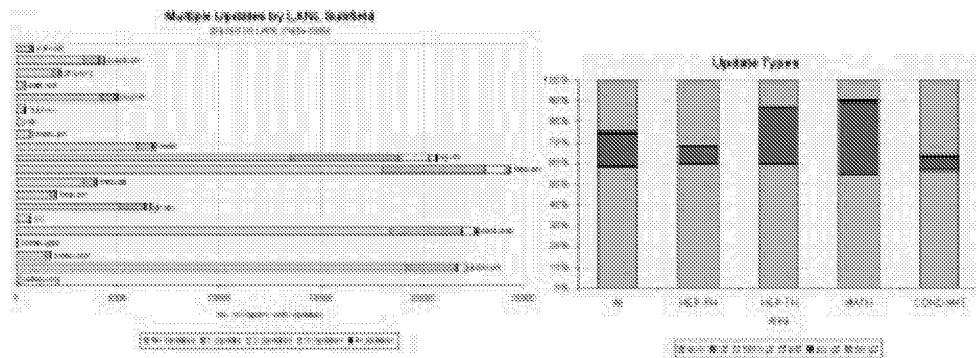


Article Embryology



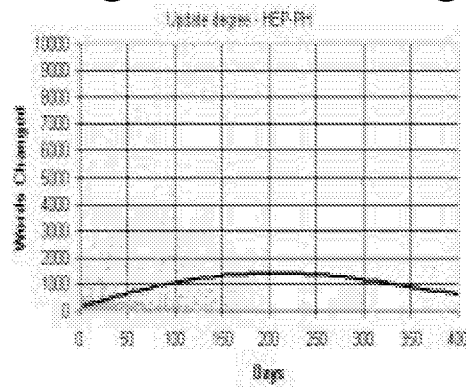
- The evolution of an article
 - Can we see evidence in the fossil record?

Versioning



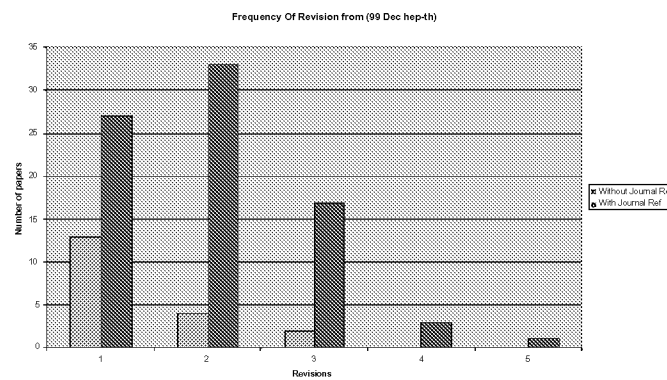
- Less versioning than we would expect

Degree of Change



- Many small changes made soon after deposit

Closeup Investigation



- First 100 articles deposited in hep-th Dec 1999
 – 81 apparently published

Closeup Investigation

- Why isn't everything published?
 - Lack of information (bad metadata)
 - Lack of quality (bad scientists)
- Why are the versions missing?
 - No versions (bad model)
 - No resubmissions (missing communication)

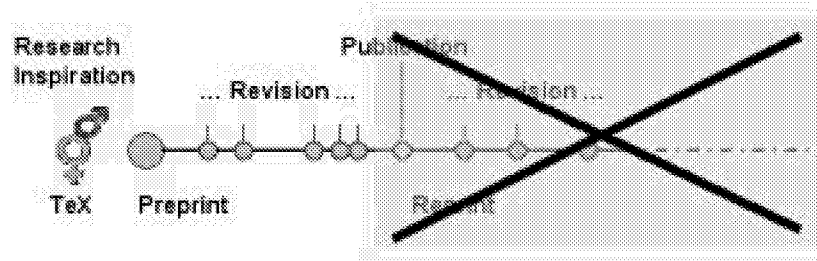


The 19 Unpublished Articles

- Initial check
 - 2 non-papers
 - 1 crackpot
 - 1 grudge
 - remaining 17 really are 'unpublished'
 - *i.e.* not journal articles
 - metadata is accurate

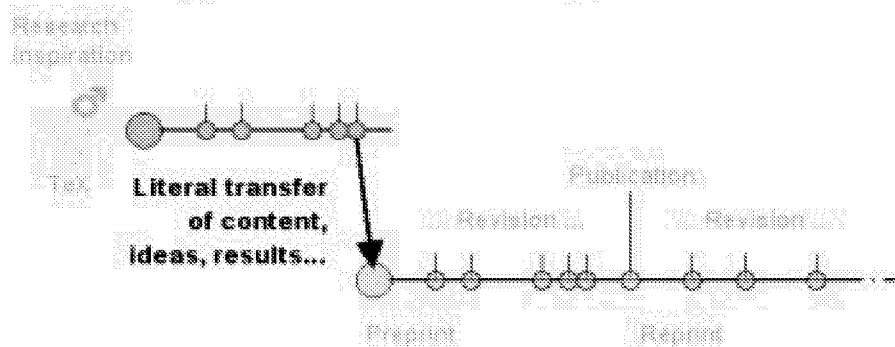


Ways of Not Being Published



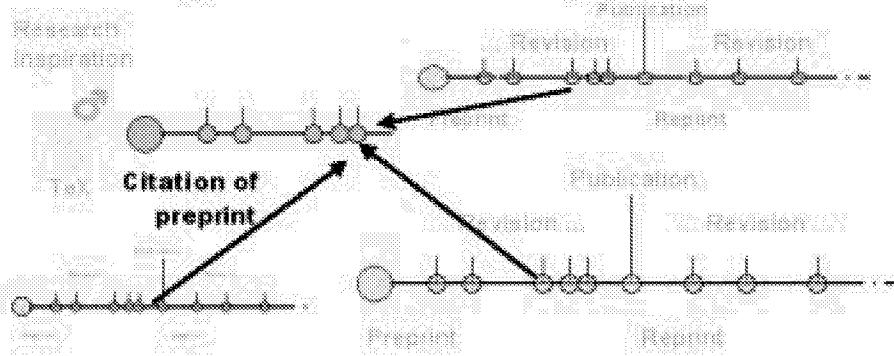
- Does non-journal status mean work is lost to the literature?
 - Dead end?
 - Uninfluential and invisible?

Ways of Not Being Published



- Not a 'dead end' if the work goes on to be published *in some other form*

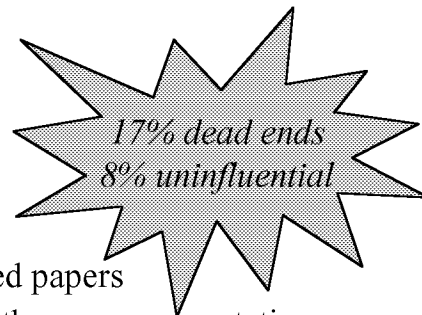
Ways of Not Being Published



- Not ‘uninfluential’ if work is cited by other publications

19 Unpublished Papers

- 2 non-papers
- 2 PhD theses / chapters
 - cited 1 / 3 times
- 3 presentations
 - all cited 4-5 times by published papers
 - 2 become new publications; other, new presentation
- 12 dead ends (*caveat emptor!*)
 - 6 cited by publications
 - 4 uncited or self-cited; previous works unpublished
 - 2 only mutually cited



Single Version Published Papers

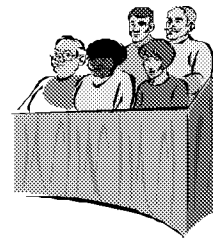
- Most likely: wait until after publication and upload final version of paper with all details
- Perhaps deposit many papers in batch

Single Version Published Papers

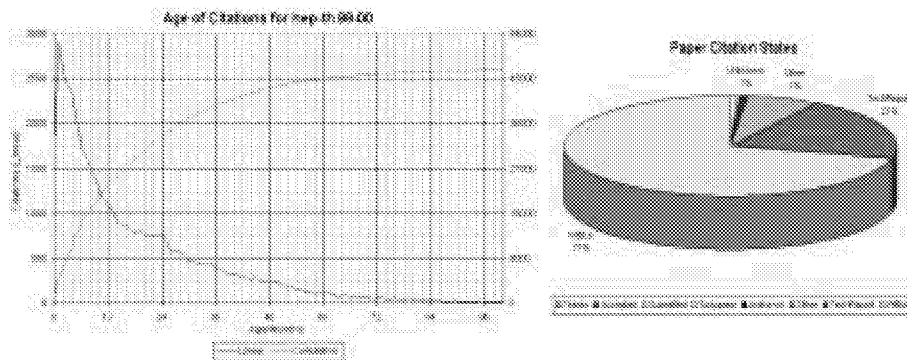
- 1 published first, deposited later
- 3 deposit and submit simultaneously
- 7 deposit first and submit later
- 7 submitted, deposited, published
- 7 couldn't trace paper
- 2 bad metadata

Judgements

- Archive is not full of unpublished and uninfluential ‘fluff’
- Metadata is pretty accurate
- Physicists are not ‘lazy’
- More investigation required!
 - Is the model right?

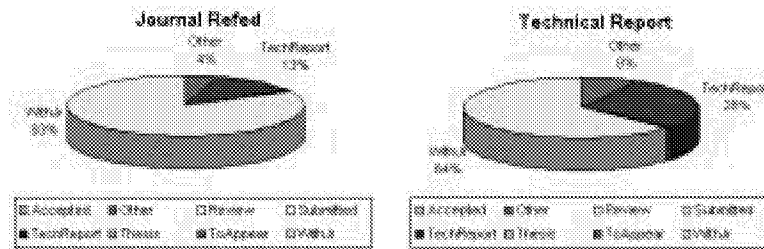


What are Authors Citing?



- Authors tend to cite
 - a lot of young articles
 - articles which *become* publications

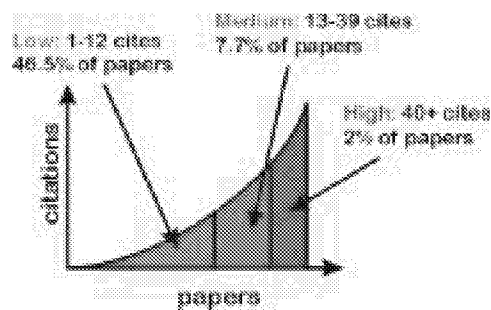
What are Authors Citing?



- Unpublished technical reports are twice as likely to cite other technical reports

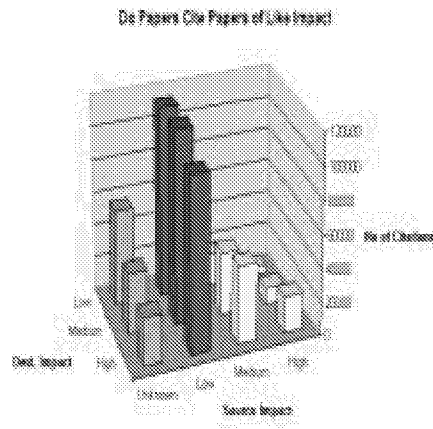
Citation Impact

| Impact | No. of Papers | Citations per paper |
|---------|---------------|---------------------|
| High | 2698 | 40+ |
| Medium | 10122 | 13 - 39 |
| Low | 61518 | 1 - 12 |
| Unknown | 57881 | 0 |



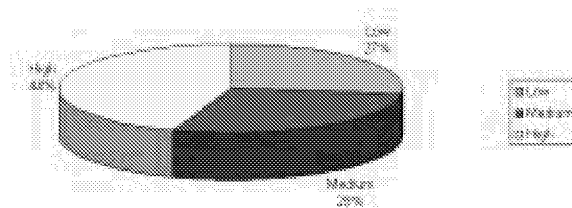
- Impact is often associated with quality

IsoCitation



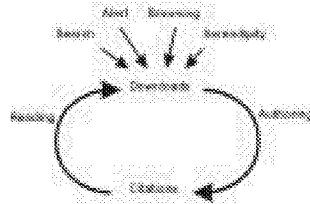
- High impact papers tend to cite other high impact papers
- Low impact papers are indiscriminate
- Medium impact papers cite medium or high impact

What Citations Are Readers Following?

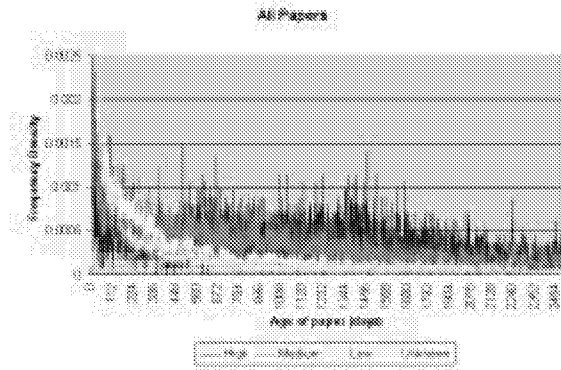


- Since citations are *by definition* evenly split, the emphasis on high impact citations is not inevitable

How Citations Affect Reading

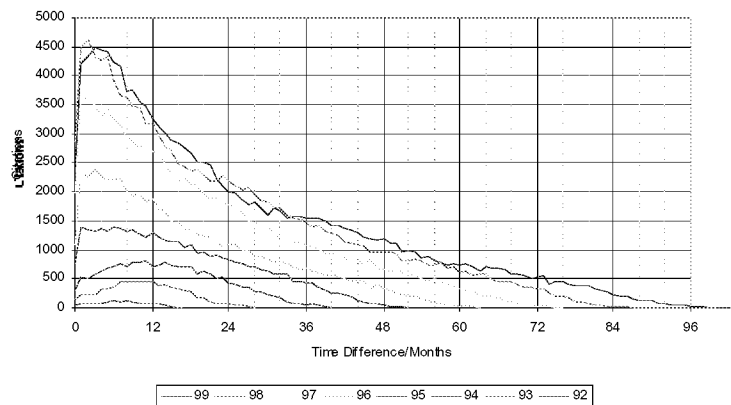


| Download type | r | n |
|----------------------|----------|-------|
| All Papers | 0.11155 | 63671 |
| High Impact Papers | 0.27293 | 1981 |
| Medium Impact Papers | 0.01288 | 5937 |
| Low Impact Papers | -0.01412 | 30163 |



- There is a correlation between highly cited papers and highly downloaded papers
- Highly cited papers have high download longevity

How Has the Archive Affected Citation Practises



- The latency of the citation peak has been reducing over the period of the archive. Speed of scientific communication is increasing. *What are the risks?*

Further Work

- Develop model of online scholarship
 - publication, research, communication
 - ask the scientist!



Questionnaire

- Questions about users' practise and attitude
- ArXiv Users 389
- ArXiv Non-Users 26
- Cogprints Users 83
- Cogprints Non-Users 166

Attitude Responses

- ArXiv users were the earliest to start and archive the earlier pre-print stages of their research
- arXiv users have lost sight of the benefits of archiving and forgotten the causal role that journals are playing in the research process.
- Cogprints users were late starters who archive the later stages of their work (post-prints)
- Cogprints users appreciate the advantages of visibility through archiving, but have unnecessary concerns over issues such as copyright

Concluding Thoughts

- Neither commentators nor users have an accurate picture of the use of archives
- What is the effect of Open Archives on the scholarly process
 - are we more effective?
 - is the literature ‘better’
- Are Open Archives a publishing mechanism for the Grid / E-Science ?

Metadata Framework for Resource Discovery of Agricultural Information

Irene Onyanchia, Fyrvola Le Hant Ward, Frehwoet Fisseha, Stefano Anibaldi, Johannes Keizer, Steve Katz
Food and Agriculture Organization (FAO) of the United Nations
Library & Documentation Systems Division (GL)
AGRIS/CARIS & FAO Documentation Group
Viale Delle Terme di Caracalla, Rome 00100, Italy
Irene.Onyanchia@fao.org, Johannes.Keizer@fao.org, Steve.Katz@fao.org

Abstract

This report outlines a proposed metadata framework for resource discovery of agricultural resources, and in particular to describe information resources in agricultural sciences. The overall work is the result of a collaborative effort between a number of partners in the agricultural community and the World Agricultural Information Centre of FAO. The endeavour is formally referred to as the "Agricultural Metadata Standards Initiative," (AG Standard). It is based upon the elements and qualifiers proposed by the Dublin Core Metadata Initiative (DCMI).

Keywords: Metadata, Interoperability, Standards, Dublin Core, Agricultural Information

1.0 Introduction

Resource description has emerged as a challenge that impedes resource discovery even though network technologies have lowered other challenges including geographical barriers. This is because resource discovery varies depending on the structure, type and content of resource and with the interests of the information keepers. Further, complex needs of users require domain specific information systems to be queried in parallel to enable access to distributed information archives. With the current enabling technology, the more complex needs of users nowadays can be met: querying more than one domain-specific information system in parallel while information managers seek to have a system that enables access to separately managed collections in-house. Example of initiatives that have been developed to encourage timely dissemination of scholarly information is the Open Archive Initiative (OAI)

To meet such demands, there is a need for a framework that would allow information access regardless of the above-mentioned barriers. The Dublin core initiative is a potential example of such a format because of its characteristics that distinguish it as a prime candidate for resource description and primary resource discovery.

The Open Archives Initiative (OAI) develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The OAI approach of interoperability attempts to combine the best of library and Internet techniques into a new model of accessing resources. It has adopted a low-barrier interoperability solution known as *metadata harvesting*, which allows content providers to expose their metadata via an open interface. The open interface prescribes to *Unqualified Dublin Core Metadata set*.

The report first provides the overall context for the metadata framework, why the standard is needed, how the work was done, and then offers thoughts on the way forward from here. Section 4 provides the elements and qualifiers of the proposed standard presented in a hierarchical structure. The hierarchical structure offers a flexible framework to implement the proposed standard at different levels of granularity, depending on the how rich each metadata source is. In its simplest form, metadata can even be supplied at the most general level of 13 core fields. A more detailed description of all the elements and the qualifiers, including information on definitions, rules, and data typing is presented in a paper at the website <http://www.fao.org/agris/MagazineArchive/MetaData/Element/1nal.doc>.

2.0 Objectives

The overall objective of the agricultural metadata framework is to define a low-barrier and richer interoperability layers using emerging standards that aim to facilitate the efficient dissemination of agricultural content. The metadata set consists of core elements and qualifiers that are generic to the description of all agricultural information resources.

The specific goals are to:

1. To facilitate the discovery of agricultural information resources;
2. To assist the management of resources by the owners;
3. To enable interoperability between different metadata structures through a simple common format;
4. To develop a metadata framework that is compliant with other standards such as MARC, ISBD and new emerging ones like Dublin Core;
5. To encode the metadata framework using new tools such as Resource Description Framework (RDF) and the Extensible Markup Language (XML). This in turn will greatly facilitate the exchange of information by:
 - Providing an overall metadata framework for better search capabilities on the internet
 - Offering a mechanism for interoperability between applications and
 - Supplying a potential means for automated processing of web resources

3.0 Strategy and Methodology Adopted

With respect to the strategy and methodology adopted to formulate the metadata framework, specific actions were taken to:

- Develop a conceptual map of the different types of information resources used in agriculture.
- Evaluate standards and common resource description practices currently used in the agricultural domain.
- Initially focus on the description of information (bibliographic) resources.
- Identify the pool of elements and qualifiers that apply to project information resources, in conformance with the guidelines of the Dublin Core Metadata Initiative.
- Develop a specific application profile for description of resources.
- Document a full element description for these resources using the set of attributes recommended by the Dublin Core Metadata Initiative.

As a result of adopting this strategy and methodology in a participatory manner with all partners of the Agricultural Metadata Standards Initiative, a list of 13 elements for agricultural resources description was proposed.

4.0 Presentation of a metadata set for the description of agricultural information resources

This section presents the proposed elements in a hierarchical structure. Preference is given to notation, vocabularies and terms that are currently used in describing agricultural resources. Full description of these elements can be viewed.

Introduction

1. The element *Creator* has been revised to represent all the agent elements namely, *Creator*, *Contributor* and *Publisher*.
2. Some attributes of elements that have been in the past considered necessary in resource description are not included in this description of a specific resource because this information is currently not considered as primary information that is important for discovery of a particular resource. However to include this information which is also important for resource discovery at a secondary level, *Authority files* shall be created and linked to the metadata. Element that will have Authority files include: *Author*, *Researcher*, *Corporate Author*, *Publisher* and *Type*, qualifiers *Event* (Conferences, Workshops, Meetings).

The hierarchical notation presents the different levels of description, which is noted by the use of different formats and colours as indicated in the legend in the footnote.

1. Proposed core elements and the qualifiers for agricultural resources description¹

Legend

| | |
|-------------------|--|
| Bold | DC Elements & Proposed elements for agricultural resources |
| <u>Underlined</u> | Qualifiers for agricultural resource description |
| <i>Italic</i> | Sub-elements or attributes of the qualifiers |
| Blue | Attributes of sub-elements |

Recommended DC qualifiers are noted with prefix (DC) while the new qualifiers are noted with (new)

1. Core element: Creator (new)

Qualifiers for Agricultural Resources

- (DC) Corporate author
Name
- (DC) Personal author
Name
- (DC) Publisher
Name
- (DC) Editor
Name
- (DC) Compiler
Name

2. Core element: Rights

Qualifiers for Agricultural Resources

- (DC) Statement
(new) Terms of use
- (new) Patent

3. Core element: Title

Qualifiers for Agricultural Resources

- (DC) Main title
- (new) Title supplement
- (DC) Alternative title

4. Core element: Relation

Qualifiers for Agricultural Resources

- (DC) Is Part Of
 - part of monograph (AM)*
(Unique identifier of related record. If URI should be PURL)
 - part of monographic series (AMS)*
(Unique identifier of related record. If URI should be PURL)
 - part of series (AS)*
(Unique identifier of related record. If URI should be PURL)
- (DC) Has Part (analytical)
(Unique identifier of related record. If URI should be PURL)
- (DC) Edition
 - Version Of Monograph,
(Unique identifier of other version)
 - URI
(Unique identifier if URI should be PURL)
- (DC) References
 - Related Records
(Unique identifiers of related record. If URI should be PURL)
 - Related Language Versions
(Unique identifiers of related language versions)

5. Core element: Subject

Qualifiers for Agricultural Resources

- (DC) Subject Classification
AGRIC Subject Categories

- CABI Codes
- LCSH
- DDC
- LCC
- UDC
- (DC) Thesaurus
 - CABI Thesaurus
 - AGROVOC Thesaurus
 - NAL Thesaurus
 - ASFA Thesaurus
- (new) Local Terms

6. Core element: Coverage

Qualifiers for Agricultural Resources

- (DC) Thesaurus
 - AGROVOC Thesaurus
 - GTN
- (DC) Country codes
 - ISO 3166 country codes

7. Core element: Description

Qualifiers for Agricultural Resources

- (DC) Abstract
- (DC) Table of contents
- (new) Notes
- (new) Holdings
 - Location
 - Address
 - Online
 - URI (PURL)

8. Core element: Date

Qualifiers for Agricultural Resources

- (DC) Publication date
- (DC) Date of Creation
- (DC) Date of Modification

9. Core element: Identifier

Qualifiers for Agricultural Resources

- URI
- ISBN
- ISSN
- Report no.
- Accession no.
- Call no.
- Patent no.
- Job no.

10. Core Element: Language

Qualifiers for Agricultural Resources

- (DC) Language of text
 - ISO 639 language codes

11. Core element: Type

Qualifiers for Agricultural Resources

- (DC) Collection
 - Monographs
 - Serials
 - Monographic serials
 - Web Pages
 - Analyticals
- (new) Document class
 - Publication
 - Non conventional
- (new) Literary indicator
 - Grey Literature
 - Legislation
 - Standard
 - Bibliography
 - Summary
 - Statistical data
 - Directory
 - Thesis
- (DC) Event
 - Conference
 - Training course
 - Workshops
 - Seminars
 - Consultations
- (DC) Images
 - Photograph
 - Film
 - Picture
 - Map
 - Slide
 - Microfiche
 - Video Cassettes
- (DC) Sound
 - Audio cassettes
 - Digital Audio
- (DC) Interactive Resources
 - CD-ROM
 - Web Forms
- (DC) Software
- (DC) Dataset
- (DC) Text

12. Core element: Format

Qualifiers for Agricultural Resources

- (DC) Medium
 - Internet Media Type, (IMT)
 - Application
 - Audio
 - Image
 - Text
 - Video
 - Print
 - Electronic media
 - CD-ROM
 - Diskette
 - DVD

(DC) Extent
 Collation
 Pagination
 Duration

1.3. Core element: Target Audience (new) Qualifiers for agricultural resources

Policy makers
 Planners
 Researchers
 Research Institutions
 Educational institutions
 Students
 Information intermediaries
 Media

Note:

The new proposed core element, Target audience will be further developed to have a standardised list. It is currently under review.

II. Full element description

This part defines each element using a set of 10 attributes recommended by DCMI that conforms to the ISO/IEC 11179 (ISO 11179) standards for describing elements. There are 10 attributes of which eight attributes were used for each element. These are *Name*, *Label*, *Definition*, *Comments*, *Language*, *Datatype* and *obligation*. The other 2 namely, *Version* and *Registration Authority* are applied globally.

The following example show how each of the elements and sub-elements was described.

Example: **Element Title**

| | |
|--------------------|---|
| Name | Title |
| Label | Title |
| Definition | A name given to the resource. Typically, a title will be a name by which the resource is formally known |
| Datatype | Alphanumeric text |
| Maximum Occurrence | Not Repeatable |
| Language | English |
| Obligation | Optional |
| Comments | |

Full description of all elements is presented at

<http://www.dublincore.org/archive/MetaData/Element/ma1.doc>

5.0 Conclusion and Future developments

A. Implementation aspects compared to generic Dublin Core

Suggestions and comments were received from all partners of the Agricultural Metadata Standards Initiative, as well as from Stuart Weibel, Executive Director of the Dublin Core Metadata Initiative. These led to the following implementation decisions with respect to the generic specification of Dublin Core:

- Merged the DC elements *Creator*, *Contributor* and *Publisher* to one main element called *Creator*;
- Dropped the element *Source*, but elaborated the element *Relation* to include information about the source;
- Proposed a new element called **TARGET AUDIENCE**;
- Proposed new qualifiers and attributes that are vital to the description and discovery of information in the agricultural domain;
- Proposed creation of *Authority files* for elements and qualifiers that have secondary information that is not included in the metadata description of a resource but is relevant for resource discovery.

The proposed metadata set for describing information resources in the domain of agriculture contain 13 elements, namely, *Creator, Title, Rights, Identifier, Relation, Description, Subject, Coverage, Date, Type, Format, Target Audience*.

B. Future developments

As mentioned earlier, this paper only represents the first step in the development of tools to aid resource discovery in the agricultural domain. The initiative will be posted and advertised in agricultural forums so as to impact the targeted audience. Work is still in progress and the logical frameworks that have been developed are in the process of being converted into technical frameworks. The proposal will also be presented to the intergovernmental process of FAO for possible endorsement by member countries.

Some of the immediate future developments are as follows:

- To encode and publish the Application Profile both as an XML Document Type Definition as well as an RDF Schema;
- To initiate a pilot project between FAO and a number of important and successful agricultural gateway services. The project aims to provide a single access point with multi-host searching using the Agricultural Application Profile as the standard for linking common metadata across the different gateway services;
- To develop software tools in support of the proposed standard (e.g. for import, export, validation, query purposes, etc.);
- To register the metadata framework and specific application profile with authoritative metadata registries;
- To develop guidelines for the application profile to assist implementers and users;
- To monitor the impact of the proposed metadata application profile for agricultural resources, making any changes or enhancements based on the results of the impact study, and undertaking outreach work to promote and facilitate the rational and widespread use of metadata.

6.0 Benefits of the application profile to FAO and the agricultural community

Format for describing and maintaining FAO in-house databases

A crosswalk of the Dublin Core, AGRIS (International Information System for the Agricultural Sciences and Technology), CARIS (Current Agricultural Research Information System) and other document repositories at FAO was developed. The crosswalk consists of the proposed core elements as container elements, while the sub-elements that qualify a specific core element are layered under the hierarchy. This mapping gives homogeneity to the different application profile under one set of defined core elements. This provides a working example of how a low level format, enables interoperability between different information systems to allow resources discovery.

Format for Data harvesting of agricultural information resources

The unqualified Dublin core based specification in its generic form supports open exchange of information initiatives such the Open Archives initiative (OAI).

Format for a unified interface for searching heterogeneous archives:- The AGRIS MHS

AGRIS Multi-Host Server is a search engine that allows parallel searching across distributed databases that are heterogeneous and have different data structure and metadata information.

The search engine is being developed in corporation with ZADI (Zentralstelle für Agrardokumentation) in Germany. It searches distributed bibliographical databases giving a one-stop access to them without the need of centralising data. The proposed application profile gives common metadata elements that homogenise search set results.

Format for resources discovery through agricultural subject gateways and information providers

Subject gateways are online services and sites that provide searchable and browsable catalogues of Internet based resources. Subject gateways will typically focus on a related set of academic subject areas. They generally consist of databases of detailed metadata or catalogued records.

Some Examples of agricultural subject gateways such as NOVAGATE, BIOME, AGRIGATE, AGNIC have the following benefits:

- Participation in a global network to bring agricultural and related information to the Web
- Offer users the opportunity to interact and resource share with other national and international agricultural institutions
- Offer opportunity to provide value-added services to constituencies

The proposed metadata frameworks offer a uniform format that could be used as a means of interoperability between these gateways. The framework offers opportunity for both low level and detailed description according to the users needs. The different levels of description indicated in the metadata framework express these.

References

1. Dublin Core Qualifiers. <http://www.dublincore.org/documents/2000/07/11/dcm-qualifiers/>
2. Dublin Core Metadata Element Set, Version 1.1. Reference description. <http://www.dublincore.org/documents/1999/07/02/dces/>
3. DMSI Type vocabulary. <http://dublincore.org/documents/dcm-type-vocabulary/>
4. Food and Agricultural Organization (FAO) Cataloguing and Indexing Manual.
5. Resource Description Framework (RDF) Model and Syntax Specification. <http://www.w3.org/TR/RDF-rdf-syntax/>
6. Metadata: standards for Retrieving WWW Documents (and Other Digitized and Non-Digitized Resources). <http://www.socsci.uci.edu/ucir/meetings/isa/fruschberg.html>
7. Interoperability Metadata Standard for Electronic Thesis and Dissertations- Version .03. <http://oldd.sspu.edu/standards/metadata/current.html>
8. Using Dublin Core. <http://dublincore.org/documents/2001/04/12/usingguide/>
9. Discovering Online Resources across the Humanities: A Practical Implementation of the Dublin Core. <http://abdn.ac.uk/publics/metadata/discovery.html>
10. Dublin Core Metadata and the Cataloguing Rules. American Library Association. Committee on Cataloguing: Task Force on Metadata and the Cataloguing Rules (Final Report). <http://www.ala.org/aleac/organization/ces/edu/01-015.html>
11. The Open Archives Initiative: Building a low-barrier interoperability framework. <http://www.openarchives.org/documents/oiu.pdf>



OAI Experiences with Arc and Kepler

<http://arc.cs.odu.edu>

<http://kepler.cs.odu.edu>

Old Dominion University

Outline

- ◆ Arc

- ◆ Kepler

- ◆ Technical Report Interchange (NASA, Los Alamos, AFRL) – Based on Arc

ARC

Arc Overview

- ◆ Arc is a federated search service based on the OAI protocol.
- ◆ By August 2001, it harvested over 1M metadata records from 38 data providers from various domains.
- ◆ Implemented experimental OAI layer over arc, supports hierarchical harvesting.

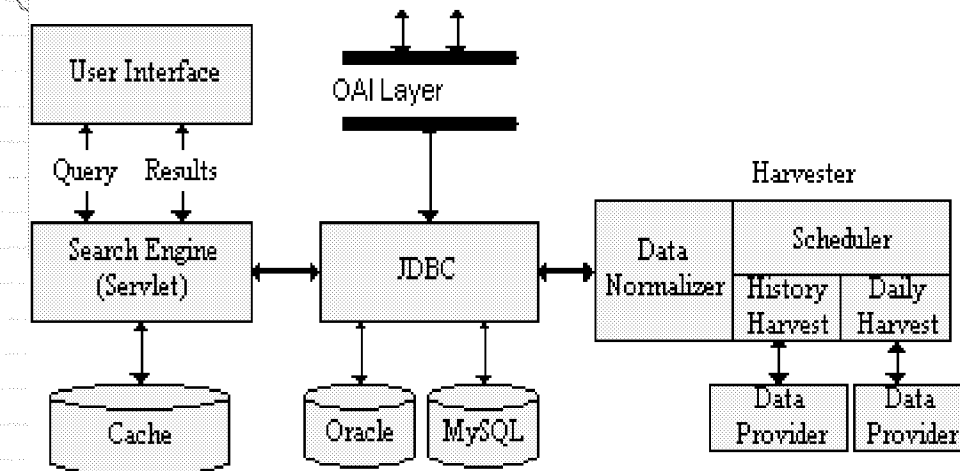
History

- ◆ Open Archive Initiative
 - <http://www.openarchives.org>
- ◆ Universal Preprint Service.
 - <http://ups.cs.odu.edu/>.
 - Initial demonstration vehicle for OAI.
 - Based on NCSTRL+ which is an extension of NCSTRL.
- ◆ Search engine developed at ODU

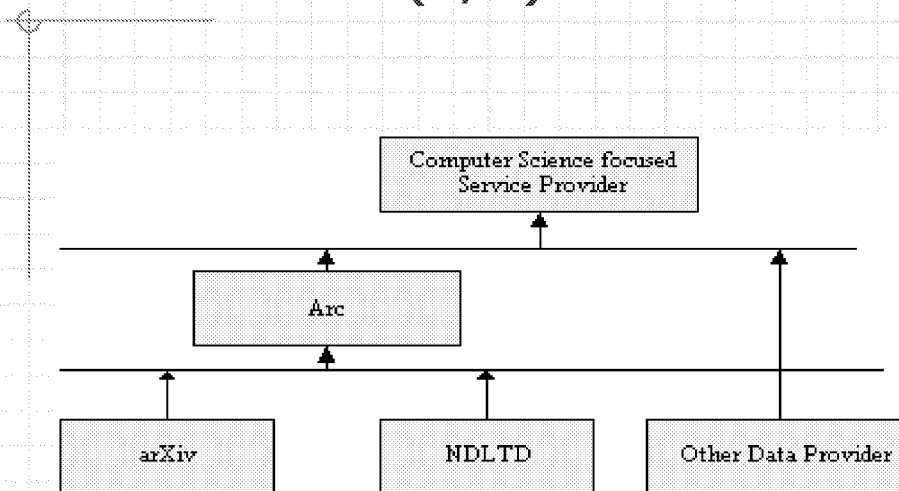
Architecture (1/3)

- ◆ Harvester.
 - Data normalization.
 - Historical harvest.
 - Fresh harvest.
- ◆ Search engine.
 - Based on Servlet/fulltext search capacity of database.
- ◆ OAI layer over arc.
 - Disseminate metadata harvested from other data providers.
 - Describe the archives from which arc harvested.

Architecture (2/3)



Architecture (3/3)



Search Service (1/2)

◆ Simple search.

- Search freetext across archive and metadata fields.

◆ Advanced search.

- Search across archives, or in specific archive and its subset.
- Search free text in author/title/abstract fields.
- Filter search/browse by archive/set/subject/type/language/datestamp/disc
overy date.

Search Service (2/2)

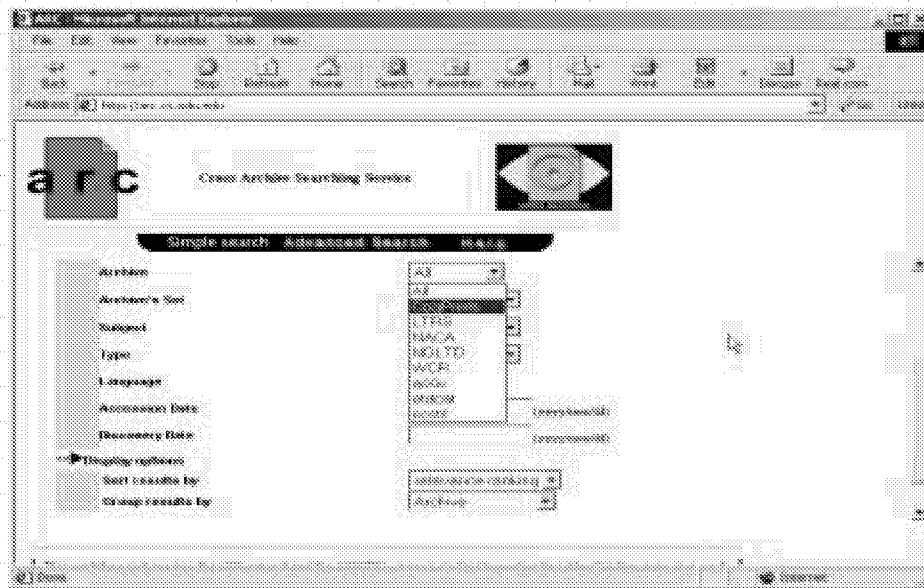
◆ Result sorting.

- By datestamp, archive, relevant ranking.

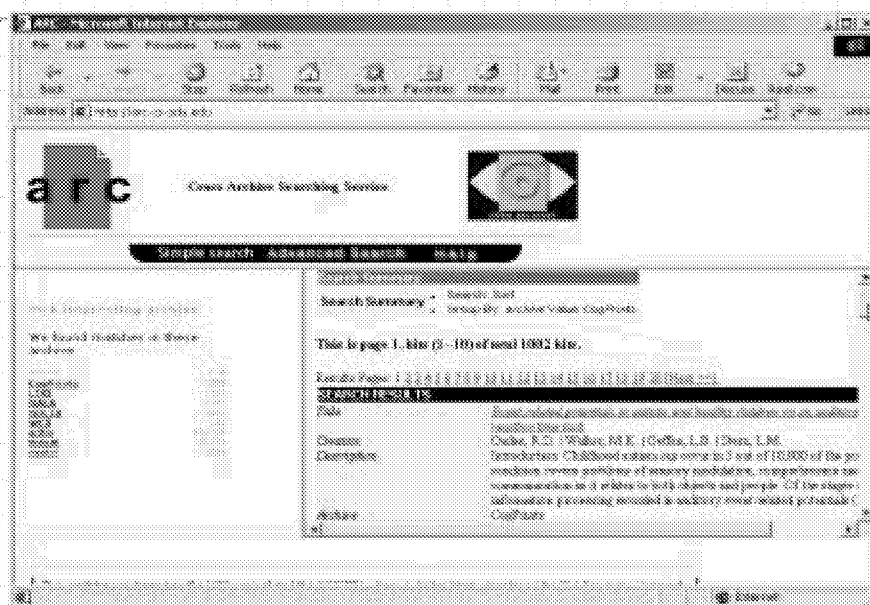
◆ Result display.

- Result list – NCSTRL+ like interface.
- Display single document in detail.
 - ◆ Lightweight bucket.
 - ◆ Link to data source.

ScreenCam- Advanced Search



ScreenCam – Search Results



Interactive Search (1/3)

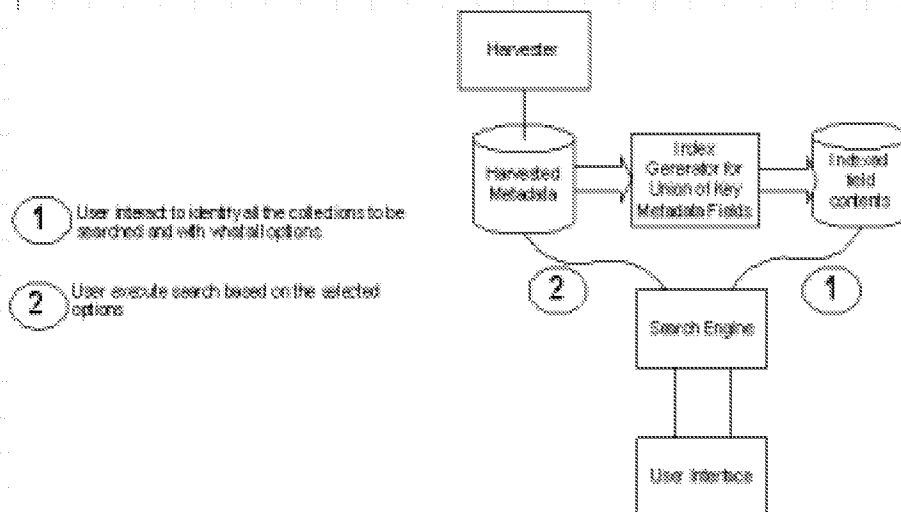
◆ Problem

- Lack of Controlled Vocabulary and Unified Search Interface

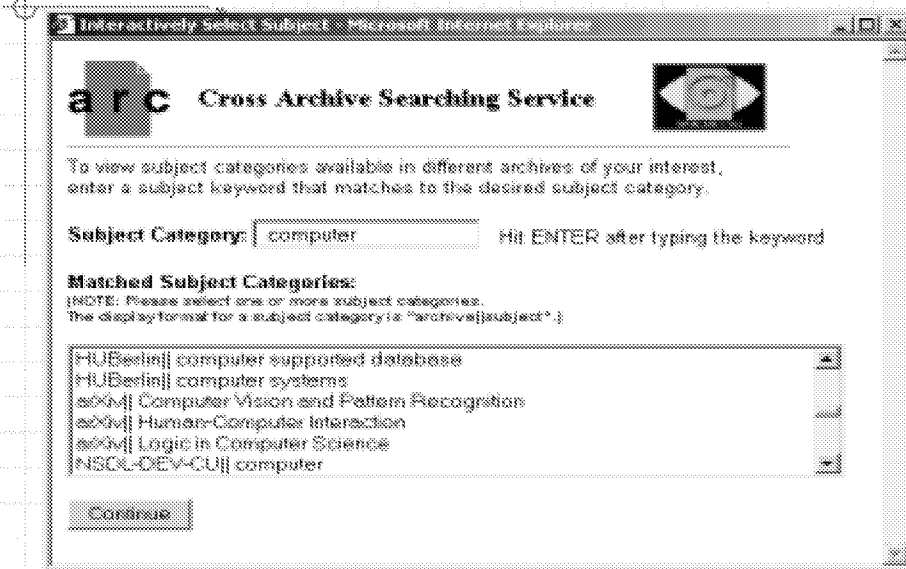
◆ Solution

- User-Centric Approach-In this approach, users have a series of interactions with the federation service to communicate their queries based on their personal model of organization of information

Interactive Search (2/3)

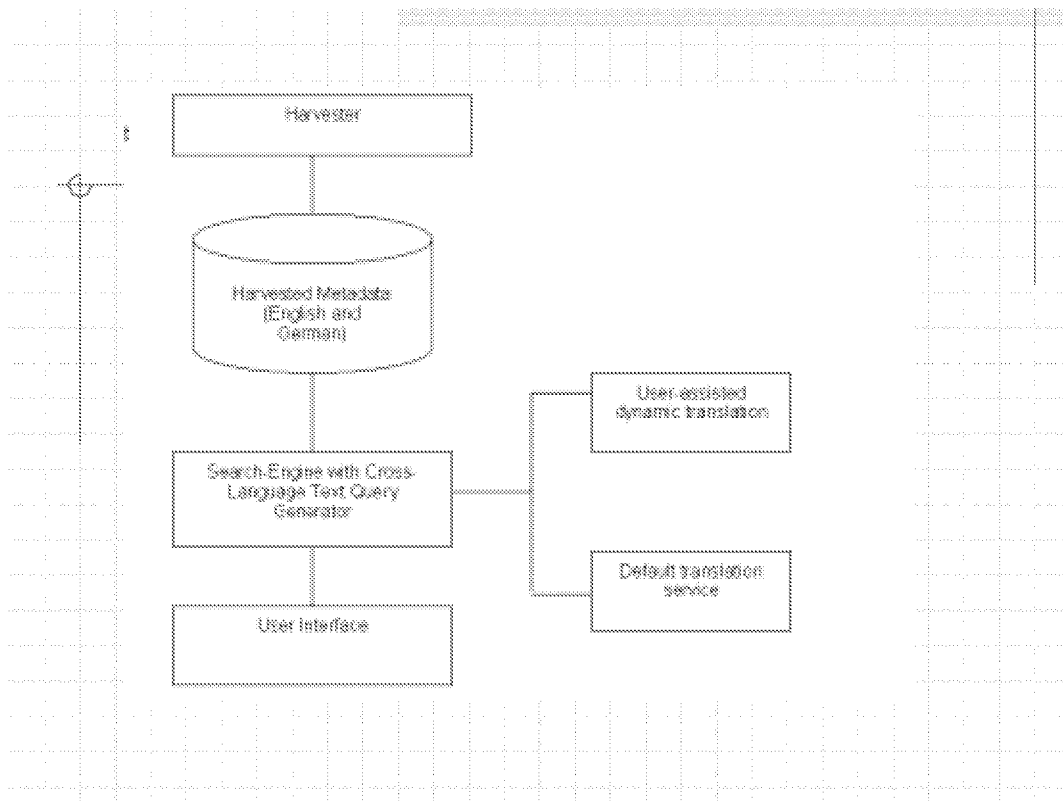


Interactive Search (3/3)



Cross Language Search

- the focus is more on how to integrate an existing cross-language technique based on query translation into Arc to support search across English and German collections.
- We maintain the harvested metadata from English collections and German collections. However, we provide a unified interface to search across collections in both languages



ScreenCam

http://130.132.4.4:8080/web/convert/search/convert-mapping.html

Your input is: project

You can choose from existing words:

Forschung

Projekt

You can input your word here:

Forschung, Projekt

Use the menus below to insert special characters:

⌘

⌘

⌘

⌘

⌘

Other

Please confirm your choice: ☒ Table ☐ Input

Submit

Close

Lessons Learned (1/2)

◆ Quality of data providers

- The expense of maintaining a quality federation service is highly dependant on quality of data providers.

◆ Controlled vocabulary

- Using unified controlled vocabulary, or at least defining mapping relationship, is important in a cross archive service.

Lessons Learned (2/2)

◆ XML syntax and character encoding

- A single error could influence large set of data.
- The character encoding error occurs frequently in most data providers.

◆ Harvest schedule

- We use historical harvest + daily based incremental harvest.
- The trade-off between data freshness and harvest efficiency.



Demo

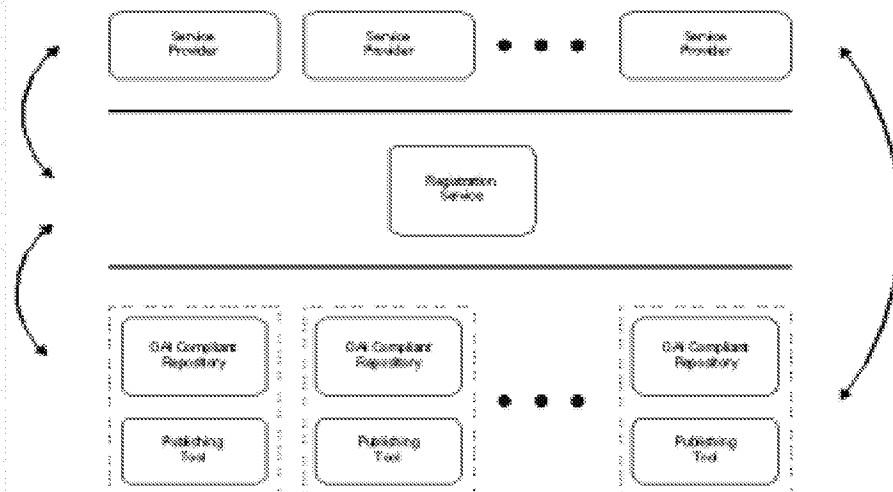


Kepler Framework

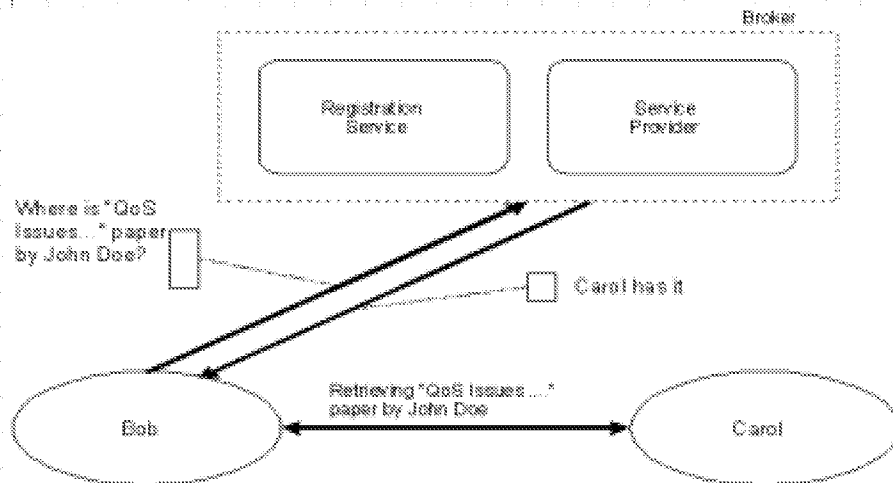
Kepler Overview

- ◆ "OAI Data/Service Provider for the Individual."
- ◆ The Kepler archivelet is available for Windows, Linux, and Unix Operating Systems. More information and download at <http://kepler.cs.odu.edu/>.

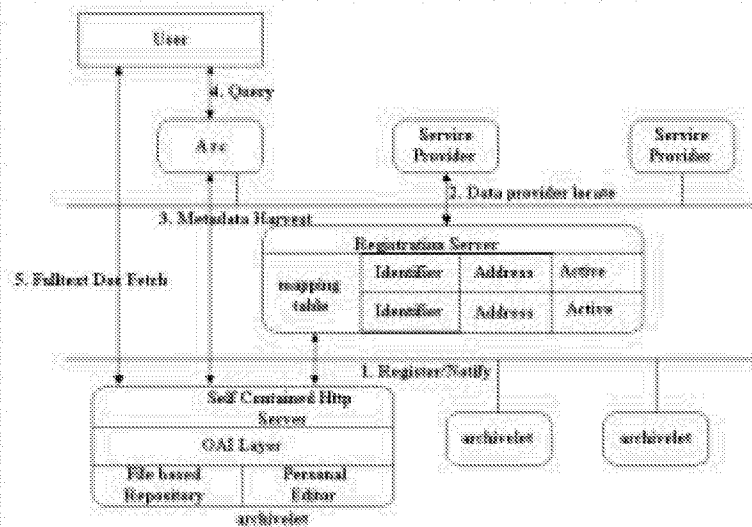
Kepler Framework



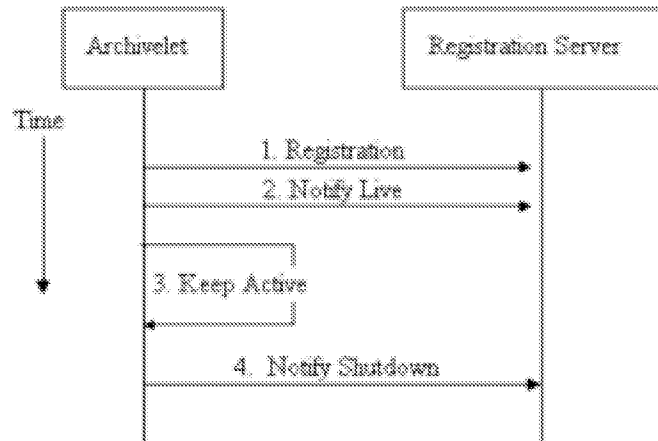
Kepler Framework and P2P model



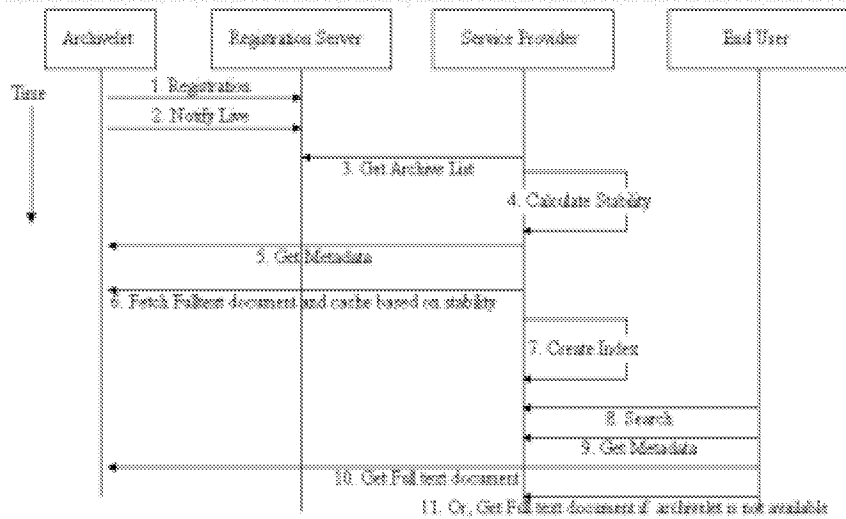
Architecture



Archivelet Registration Process



Kepler Process



Archivelet Interface

Edit user profile

User Status Information

| | |
|------------------------|------------------------|
| Repository identifier: | acorn:100 |
| URL: | http://128.82.7.74:201 |
| Registered: | Y15 |
| Online: | Online |

Edit User Profile

| | |
|----------|----------------------|
| name: | warning |
| content: | personal publication |
| email: | ru_w@cs.umd.edu |

OK Cancel

Edit Metadata

Source: H:\epaper\din\olap\olap1.pdf

Browse

Metadata Information

| | | | |
|-------------|------------------------|-------|------------|
| Identifier: | 9-101 | | |
| Title: | Peer to Peer Computing | | |
| Creator: | gubert | Edit | |
| Format: | pdf | date: | 2000-05-09 |
| Subject: | computer science | Edit | |
| Language: | en-us | | |

Description

which aim to make digital libraries (DLs) easier to use, both for publishers and retrievers. We are interested in DL tools, DL interoperability, and DL architecture. Our research is supported by the NSF and NASA. We especially work closely with NASA Langley Research Center, the Lead Center for Scientific and Technical Information in NASA.

More

OK Cancel

Kepler - slaoming's Achieve

Add Edit Remove Status

| Identifier | Title | Creator | Date |
|------------|--------------------------------|------------|------------|
| tr-100 | Learning Perl | M.L.Nelson | 2001-01-01 |
| tr-101 | Peer to Peer Computing | zubair | 2000-09-09 |
| goodlucke | Applied XML | Naidu | 2000-09-09 |
| tr-999 | test | le | 2000-09-09 |
| tr-102 | Webster's new World Dictionary | Newton | 1999-01-01 |
| tr-103 | XML specification guide | zubair | 2000-09-09 |
| tr-104 | Build XML applications | Booch | 2000-09-09 |
| slaoming | this is a test for slaoming | tom | 2000-01-01 |
| mm | mm | mm | 2001-01-01 |

View Close

Kepler Search Service

Kepler Search Service

Simple search:

Full text search window

Not found matches in these archive

2000-09-09-09
2000-09-09-09
2000-09-09-09
2000-09-09-09
2000-09-09-09
2000-09-09-09
2000-09-09-09

Search Summary:

Group By: archive date (1/1/2000-9/9/2000)

Search Results

Title: Integration of open literature and electronic journals in the OER

Comment: Personal Contact

Description: The presentation covers three points: the OER implementation at the Library, the current OER project processing procedure and using to manage the usage of the journals with a new service.

Archive: 2000-09-09

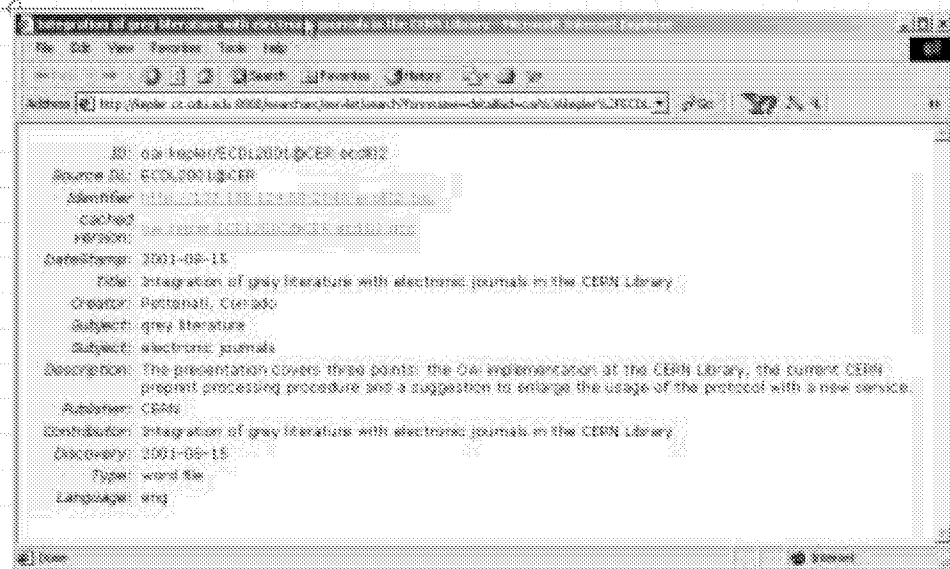
Date Range: 2001-01-01

Document ID: 2000-09-09-09-09-09-09-09-09

Title: Integration of open literature and electronic journals in the OER

Abstract

Document with cached version



Issues

- ◆ Acceptability of the individual focus framework
- ◆ Accessibility of archivelets all the time
- ◆ Registration Service: It's important for Kepler, and OAI community also needs it with the increment of number of OAI DPs
- ◆ What's the potential benefit of Kepler Model comparing with web-based publication?



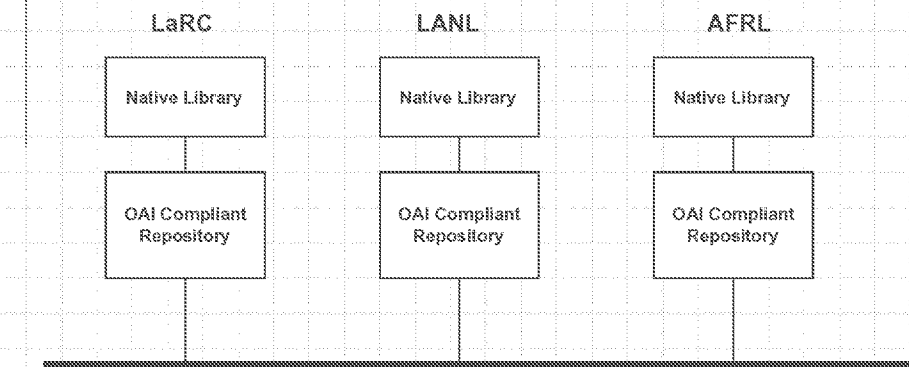
Demo



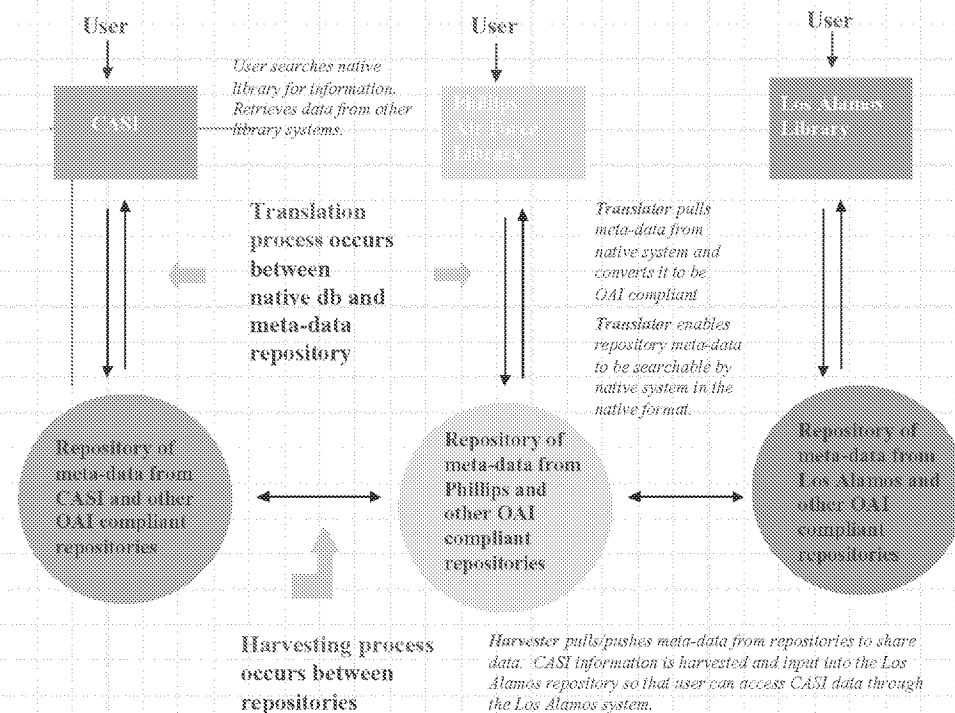
Technical Report Interchange (TRI)

*In Cooperation With NASA Langley Research Center (LaRC),
Los Alamos National Laboratory (LANL), and the Air Force Research
Laboratory (AFRL)*

TRI Objective

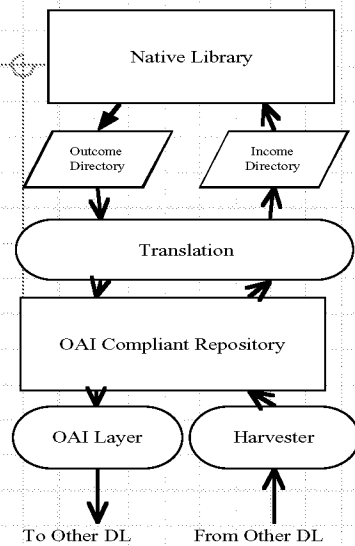


TRI Project NASA/Air Force/DOE R&D by ODU



By JoAnne Rucker, NASA Langley Research Center

A Typical TRI Module -- Phase I



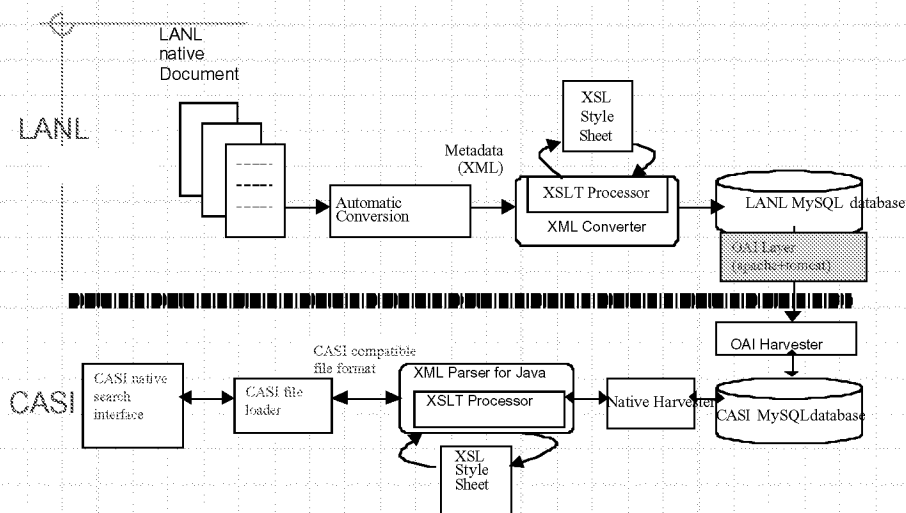
•XML based repository with OAI layer

•Tool to populate the repository from other OAI compliant archives

•Tools to integrate the XML repository with the native library

LANL

A typical workflow- CASI shares documents created by LANL



Finished work

- ◆ OAI Harvester
- ◆ OAI layer over MySQL/Oracle database
- ◆ Native Harvester
- ◆ Translation Module for LANL
- ◆ Scheduler which coordinates all local management work and keeps log file.

TRI- In Progress Work

- ◆ Translation module for CASI and AFRL
- ◆ Subject Mapping
- ◆ Explore the possibility of definition of GTRM (Government Technical Report Metadata) XML Schema.
 - To exploit the rich metadata schema beyond Dublin Core

LANL Native metadata format

```
<id>00418048</id>
<url>http://www.lanl.gov/la-pubs/00418048.pdf</url>
<author>Los Alamos National Laboratory</author>
<title>Computation of the return current in encephalography : The auto solid
angle</title>
<crefnum>LA-UR-90-1947</crefnum>
<crefnum>DE90013186</crefnum>
<crefnum>CONF-900756-11</crefnum>
<controlnum>{OSTI88} 98000276065</controlnum>
<controlnum>NndLa5-DOE12370019</controlnum>
<publisher>Los Alamos, NM : Los Alamos National Laboratory,</publisher>
<pubDate>[1990]</pubDate>
<subClass>UC--405</subClass>
<subClass>990200</subClass>
<subClass>550300</subClass>
<description>15 p.</description>
<indexing>Sponsoring organization acronym: DOE/ER W-7405-ENG-36</indexing>
<abstract>In the integral equations that arise in potential theory it is necessary to
evaluate the solid angle subtended at a point r on the surface by each element of
area of the surface. If the surface is approximated by a set of triangles then those
triangles than contain the point r as a vertex require special attention, and this is
referred to as the "auto solid angle" problem. We give an exact formula for
evaluating the auto solid angle, and a prescription for apportioning it amongst the
vertices of these triangles. 5 refs., 4 figs.</abstract>
<notes>Conference literature</notes>
<Page>
```

Mapping Table

[illegible]

LANL Report in OAI-DC format

```
<record>
<recordInfo>
<identifier>id:TRI:lanl/00418040</identifier>
<datestamp>2001-07-29</datestamp>
</recordInfo>
<metadata>
<dc:source>http://purl.org/dc/elements/1.1/
<dc:source>http://www.w3.org/2000/10/XMLSchema-instance
<dc:schemaLocation>http://purl.org/dc/elements/1.1/
http://www.openarchives.org/OAI/DC.xsd
<dc:title>Computation of the return current in encephalography : The auto solid
angle</dc:title>
<dc:creator>Los Alamos National Laboratory</dc:creator>
<dc:subject>LC--4815 eddneur</dc:subject>
<dc:subject>990200 eddsc</dc:subject>
<dc:subject>550300 eddsc</dc:subject>
<dc:publisher>Los Alamos, NM : Los Alamos National Laboratory,</dc:publisher>
<dc:contributor>Computation of the return current in encephalography : The auto solid
angle</dc:contributor>
<dc:date>[1990]</dc:date>
<dc:type>Conference Literature</dc:type>
<dc:identifier>http://lb-www.lanl.gov/la-pebs/00418040.pdf</dc:identifier>
<dc:source>LANL</dc:source>
<dc:language>English</dc:language>
<dc:relation>LA-UR-90-1947|DE90013186 O5II|CONF-900756-11*(O5TIED08)
90000276065|(Nml.e5-DGE)12370019</dc:relation>
</dc:metadata>
</record>
```

CASI Metadata File format

TRI

```
<0000>[000]>1990011992
<0010>[000]>No Copyright
<0030>[000]>0030<001>>CASI:001|Videotape-Data
<0030>[000]>0030<001>>CASI:001|Videotape-VHS
<0030>[000]>00000000
<0030>[000]>19901117
<0040>[001]>English
<0070>[001]>NASA Scope and Coverage
<0070>[001]>16:Space Transportation
<0080>[001]>NASA-SCS-VT-19900700
<0090>[000]>N
<0140>[000]>STB-410 TCIT
<0140>[000]>19940015
<0160>[000]>Sep. 15, 1994
<0160>[000]>1994
<0300>[001]>Videotape: 2D min. playing time, in color, with sound
<0350>[000]>Unclassified
<0350>[000]>NASA
<0350>[000]>Unrestricted - Publicly Available
<0350>[000]>NASA
<0500>[001]>Technical Report
<0500>[001]>The Case of STB-410 Challenger Shuttle, Pilot Jon A. McBride,
e. Mission Specialists Michael J. Smith, S. Sullivan, Sally K. Ride and David C. Lee,
and Payload Specialists Marc Garneau, and Paul B. Scobee-Power are seen driv-
ing in the Shuttle to pick up the Commander of the mission Robert L. Crippen.
Footage of the crew exiting at the launch pad, departing the Shuttle-VSS and
boarding the shuttle to perform a trial countdown demonstration test are shown.
Members of the Challenger team are seen exiting the shuttle, and answering of
questions from reporters. Live footage of the flight control room, and several
panoramic views of the shuttle on the launch pad are also seen.}
<0520>[000]>NC
<0520>[000]>CASI
<0540>[000]>Regular
<0540>[000]>Videotape-Data
<0540>[001]>CHALLENGER
<0540>[001]>PREFLIGHT TESTS
<0540>[001]>PRELIMINARY TESTS
<0540>[001]>PRELIMINARY TESTS
<0540>[001]>TEXT FLIGHT
<0540>[001]>SPACE SHUTTLE MISSION 41-0
<0540>[001]>CHALLENGER (ORBITER)
```

CASI Mapping table

(available from oai.larc.nasa.gov)

OAI Summary Record: MARC-to-Dublin Core Mapping
(grouped by DC field)

Page 45 of 45

230. *Staphylinus* 3 more specimens

| HEC# | Tag | Field Name | Boolean Core Mapping Note |
|------|-----|---------------------------|--|
| 245a | 75a | Including Scientific | Corresponds to tag field for following fields in the order shown:
245a, 245b, 245c, 245d, 245e, 245f |
| 245b | 75b | Public use | Represents title field with scientific field along with 245a, "P" flag that has no "S" will be ignored in a public use tag |
| 245c | 75c | Department | See 245a |
| 245d | 75d | Department Number of Part | See 245a |
| 245e | 75e | Department Name of Part | See 245a |
| 245f | 75f | Department Version | See 245a |

1502-2756

| Is 183 Tag | Field Name | Default Value Mapping |
|------------|------------|-----------------------|
| | | |

CASI report in OAI-DC format

X-ray crystallography

<http://www.port.org/de/dokument/7.1.1/> <http://www.port.org/2008/7/1/2008-06-09-berlin>

<http://port.org/de/dokument/7.1.1/>

<http://www.gutenberg.org/files/14314/1-4>

Seitliche Cammeration und Horizont: Outgassing Standalone – An Interactive Helium Helium

© 2000 Blackwell Science Ltd *Journal of Internal Medicine* 247: 395–402

The goal of this program is to collect at one site some of the knowledge accumulated about the outgassing properties of aerospace materials based on ground testing, the effects of this outgassing observed on spacecraft in flight, and the in-orbit contamination environment measured by instruments on orbit. We believe that this work will help some contamination in the future, and that the information collected will be useful to the aerospace community. The program will be carried out in a number of phases. The first phase will be to collect information from the literature, from the knowledge base, and also from some information from ground testing and space missions, regarding and summarizing the state of the art in the major contamination and outgassing areas to indicate that information and ground testing, major spacecraft design, present and future spacecraft, and ground defense decisions.

Downloaded from <http://ajphaphapublications.sagepub.com/> at National Archive Publishing Co on June 11, 2015

03481X-AN9848

● 2008 年 12 月 1 日起, 凡在境内销售货物或提供应税劳务, 以及进口货物, 均须向税务机关申报并缴纳税款。

52001118 **RECENT CONCEPTS IN MATERIALS**

5393 313077 (C038500)2322

[illegible][illegible]

www.pearsoned.com.au

COPYRIGHT © 1987 BY

[illegible]

Green, D. B. 1983.

© 1997 by Physical Sciences, Inc., Redwood, CA, United States

1001-8381

Abstract

1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 26

TRI Future Directions

- ◆ Integration of Other Native Libraries
- ◆ Integration of search service --
Federation local libraries (Arc)
- ◆ Publishing tools for TRI repository
- ◆ Extend OAI
 - Channel based subscription for auto
synchronization

Demo

Conclusion

OAI is making feasible to build higher level services that federates metadata from heterogeneous resources and provide a unified interface for users to access these resources.

Open Archive Forum

European Support for Open Archives

Susanne Dobratz
Humboldt-University Berlin
Computing Centre
dobratz@rz.hu-berlin.de

Open Archive Forum

Contents

- Project Background
- Project Objectives
- Practical Details

Open Archive Forum

Partners

- European Union Information Society Technologies (IST) programme accompanying measure
- Project start October, 1st, 2001 (2 years)
- Partners:
 - UKOLN, University of Bath
 - IEI-CNR, Pisa
 - Humboldt-University, Berlin

Open Archive Forum

OAI and OAForum

- OAIs main objective:
(according to Carl Lagoze)
 - develop a protocol + promotion of that
 - final version expected in April 2002
 - Workshops in April/May 2001 in Washington and Europe (Pisa)
 - currently provides a registration service for OAI data providers and service providers

Open Archive Forum

Background

- **UKOLN**
 - various projects contexts of metadata and interoperability, cross searching
 - Renardus, Schemas, DESIRE
- **CNR**
 - CYCLADES project, DELOS
 - develop services on top of OAI specification
- **Humboldt-University**
 - Dissertationen Online
 - DINI Workshops

ECDL2001, 08.09.2001

(c) Susanne Dobratz, Humboldt-University

5

Open Archive Forum

Objectives

- **building and supporting communities**
 - support projects, national initiatives
 - encourage them to sharing experiences
 - sharing sharing software, tools, metadata standards
- **facilitate a critical evaluation of the potential of OAI**
 - validating European experiences with OAI specification

ECDL2001, 08.09.2001

(c) Susanne Dobratz, Humboldt-University

6

Open Archive Forum

Objectives

- interoperability
 - OAI vision: deploy a technology giving
 - a low cost
 - low entry barrier to interoperability
 - scalability of distributed searching
- metadata schemas
 - OAI allows additional parallel use of metadata schemas
 - explore further possibilities

ECDL2001, 08.09.2001

(c) Susanne Dobratz, Humboldt-University

7

Open Archive Forum

Objectives

- Variety of content and business models
 - potential for new services
 - exploit commercial opportunities
 - leads to more consequences
 - help European stakeholders to form a coherent framework for future policy decisions
 - provide access to a broad range of digital resources (e.g. digitized materials)

ECDL2001, 08.09.2001

(c) Susanne Dobratz, Humboldt-University

8

Open Archive Forum

Interest of European Union

- Release value of the invisible web
- Low cost interoperability
- Influence developments, ensure European perspective
- Advantages of clustering projects
 - collaboration
 - exchange of information

Open Archive Forum

Partners

- Who will be involved?
 - Cultural heritage institutions
 - Research organisations
 - Public library sector
 - Community services
 - Commercial sector
 - Education sector
- As service provider:
 - E-print services
 - Aggregators
 - Value adding services
- As data provider:
 - Establish metadata repositories
 - New data providers

Open Archive Forum

Explore Content Models, Business Models

- Cooperation of data providers to form a network of service providers
 - used metadata schematas
 - undefined pieces in protocol (about-container, set definitions)
- Data providers will be service providers

Open Archive Forum

Explore Content Models, Business Models

- Metadata provided „free“ to service providers
 - services not free
- Provision of value-added services
 - document delivery
 - services for targeted audiences
- Investigations:
 - Outcome: expert reviews of key issues

Open Archive Forum

Evaluate OAI protocol

- Compare with existing technologies
 - Z39.50, Harvest
- Dublin Core used
 - Qualifier needed?
 - Further community specific standardisation needed?
- What type of organisations could best exploit Open Archives
- Which are the benefits to the users?

ECDL2001, 08.09.2001

(c) Susanne Dobratz, Humboldt-University

13

Open Archive Forum

Encourage broad usage

- Workshops
 - series of four
 - informed by domain specific reports
- OAI EU Information Source
 - Inventory of software tools
 - Interoperability issue register
 - Current implementation database
 - additional information / valuations / Comments

ECDL2001, 08.09.2001

(c) Susanne Dobratz, Humboldt-University

14

Open Archive Forum

Please register interest!

- Preliminary Website:
 - <http://edoc.hu-berlin.de/oaf>
- Contact:
 - Rachel Heery, Leona Carpenter
r.heery@ukoln.ac.uk,
l.carpenter@ukoln.ac.uk
 - Donatella Castelli
castelli@iei.pi.cnr.it
 - Susanne Dobratz:
dobratz@rz.hu-berlin.de

ECDL2001, 08.09.2001

(c) Susanne Dobratz, Humboldt-University

15

*European Conference on Digital Libraries (ECDL)
Darmstadt, 8 September 2001*

***Experimental OAI Based Digital
Library Systems -
Workshop for OAI Implementers***

Closing Remarks

Workshop Results

- Conflict between Heterogeneous and "Homogenized" Metadata to Achieve Semantic Precision &
- Uniform, Valid Metadata (Field content, Format, Controlled Vocabularies, Community Standards ...)
- Metadata extraction from existing "trusted" metadata holdings (with augmentation, editing ...) vs. / combined with Author-produced Metadata
- Unified Search Interface vs. User-Customized ... User-Centric- / User-Driven-Approach
- Analysis of Use and Submission-Publication Relationships, as well as of citation / references of OAI repository documents

Workshop Results (2)

- Cross-language Searching / Multilinguality
- Multi-language Metadata / User-assisted Dynamic Translation
- Hurdles in Cross-Archive Harvesting due to Community demands on additional Metadata
- Archivelet Registration / Creating a small-scale OAI Server
- Political Issues – Institutional Archival Servers, Convincing the Stakeholders While Assuring Scholarly Prestige & Evaluation Basis
- Ways and Means to Join the European OA-Forum

What We Didn't Talk About

(Or only a Little!)

- Implementation Assistance
- Metadata harvesting for Multimedia (non-document) items

Proceedings and Action Points

- Proceedings prepared with Kepler
 - Presenters submit their presentations!
 - Implementers register their repository!
 - Take comments, suggestions and unanswered questions to OAI-Tech (Zubair, Mike), DC-CITE (Diann)
 - Notification when Proceedings complete
 - Feedback and later comments on Workshop
-

Prof. Mohammad Zubair
zubair@cs.cdu.edu

Dr. Diann Rusch-Feja
ruschfeja@mpib-berlin.mpg.de

| REPORT DOCUMENTATION PAGE | | | Form Approved
OMB No. 0704-0188 | |
|---|---|--|-------------------------------------|----------------|
| Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503. | | | | |
| 1. AGENCY USE ONLY (Leave blank) | 2. REPORT DATE
April 2002 | 3. REPORT TYPE AND DATES COVERED
Technical Memorandum | | |
| 4. TITLE AND SUBTITLE
Experimental OAI-Based Digital Library Systems | | 5. FUNDING NUMBERS

992-16-05-02 | | |
| 6. AUTHOR(S)
Michael L. Nelson, Kurt Maly, Mohammad Zubair, and Diann Rusch-Feja (Editors) | | | | |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

NASA Langley Research Center
Hampton, VA 23681-2199 | | 8. PERFORMING ORGANIZATION
REPORT NUMBER

L-18186 | | |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Washington, DC 20546-0001 | | 10. SPONSORING/MONITORING
AGENCY REPORT NUMBER

NASA/TM-2002-211638 | | |
| 11. SUPPLEMENTARY NOTES
Primarily viewgraphs. | | | | |
| 12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified-Unlimited
Subject Category 82 Distribution: Standard
Availability: NASA CASI (301) 621-0390 | | 12b. DISTRIBUTION CODE | | |
| 13. ABSTRACT (Maximum 200 words)
The objective of Open Archives Initiative (OAI) is to develop a simple, lightweight framework to facilitate the discovery of content in distributed archives (http://www.openarchives.org). The focus of the workshop held at the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001) was to bring researchers in the area of digital libraries who are building OAI based systems so as to share their experiences, problems they are facing, and approaches they are taking to address them. The workshop consisted of invited talks from well-established researchers working in building OAI based digital library system along with short paper presentations. | | | | |
| 14. SUBJECT TERMS
OAI, Open Archives Initiative, Digital Libraries, Interoperability, Metadata Harvesting | | | 15. NUMBER OF PAGES
140 | 16. PRICE CODE |
| | | | | |
| 17. SECURITY CLASSIFICATION
OF REPORT
Unclassified | 18. SECURITY CLASSIFICATION
OF THIS PAGE
Unclassified | 19. SECURITY CLASSIFICATION
OF ABSTRACT
Unclassified | 20. LIMITATION
OF ABSTRACT
UL | |